SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers

JOEL CHAN, University of Maryland
JOSEPH CHEE CHANG, Carnegie Mellon University
TOM HOPE and DAFNA SHAHAF, Hebrew University of Jerusalem
ANIKET KITTUR, Carnegie Mellon University

Scientific discoveries are often driven by finding analogies in distant domains, but the growing number of papers makes it difficult to find relevant ideas in a single discipline, let alone distant analogies in other domains. To provide computational support for finding analogies across domains, we introduce Solvent, a mixed-initiative system where humans annotate aspects of research papers that denote their background (the high-level problems being addressed), purpose (the specific problems being addressed), mechanism (how they achieved their purpose), and findings (what they learned/achieved), and a computational model constructs a semantic representation from these annotations that can be used to find analogies among the research papers. We demonstrate that this system finds more analogies than baseline information-retrieval approaches; that annotators and annotations can generalize beyond domain; and that the resulting analogies found are useful to experts. These results demonstrate a novel path towards computationally supported knowledge sharing in research communities.[1]

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**;

Additional Key Words and Phrases: Scientific discovery; computer-supported cooperative work; analogy; crowdsourcing

ACM Reference format:
Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 31 (November 2018), 21 pages.
https://doi.org/10.1145/3274300

1 INTRODUCTION

Analogies are an important driver of scientific progress. For example, Darwin's theory of evolution was conceived by analogy to human population dynamics [21]; Salvador Luria's Nobel-Prize-winning work on bacterial mutation was inspired by analogy to a slot machine[31]; the simulated annealing optimization algorithm was inspired by the annealing process for removing imperfections in metals [26]; and information foraging theory was inspired by analogy to how animals forage for food [36, 41].

[1]This abstract demonstrates the annotation scheme we use in our mixed-initiative system: background (yellow), purpose (red), mechanism (blue), and findings (gray).

2573-0142/2018/11-ART31 $15.00
https://doi.org/10.1145/3274300

However, as the number of papers grows, individual researchers increasingly struggle to discover relevant analogies from work within their own discipline, let alone analogies across different domains or disciplines (e.g., from metallurgy to optimization, for simulated annealing). The problem is especially acute as many fields of study now require recombining knowledge across multiple disciplines [24, 45]. For example, cognitive science relies on insights from linguistics, artificial intelligence, neuroscience, cognitive psychology, and education; creativity and innovation research must integrate insights from the study of individual cognitive factors, motivational/emotional factors, team dynamics, and societal/economic incentive structures [38]. Individual scientists lack the time and resources to keep up with all the possible conferences, journals, and talks that might hold analogical insights for problems they are working on. For these reasons, computational systems that can mine analogies between research papers could significantly accelerate scientific progress by reducing the cost of knowledge discovery.

A core challenge to building such systems is creating appropriate semantic representations of documents that can support analogical matching. The essence of analogy is matching a target knowledge representation (e.g., document, abstract, research problem description) with other source knowledge representations that share its core relational structure [15]; when the analogous sources also have many other details that are very different from the target, they are known as far or domain-distant analogies. For example, the relational structure of the annealing process can be described as "MINIMIZING imperfections in a substance BY HEATING and gradually COOLING the substance" (relations in ALL-CAPS). This relational structure can then be analogically matched to other sources that share all or part of the structure, such as optimization problems, which share a similar relation of "MINIMIZING error in a model", despite vast differences in their domain features (e.g., physical temperature and metals VS. algorithms and digital data).

We build on promising recent work by Hope et al. [22] that takes a mixed-initiative approach to creating such representations for consumer product descriptions: 1) crowds annotate aspects of the documents that denote their ***purpose*** (what they are trying to achieve) and ***mechanism*** (how they achieve that purpose), and 2) these annotations are used to construct semantic vector representations that capture the overall purpose and mechanism of each document. The intuition behind this approach is that the overall representations of purpose and mechanism constitute a "soft" relational *schema* [16] because of the in-built causal relation between purpose and mechanism. This soft schema can then be used for analogical queries (e.g., finding other products with similar purpose but different mechanisms – a classic analogical problem solving query [16]). Hope et al. demonstrated that crowd workers could produce purpose and mechanism annotations reliably and quickly (usually in under a minute per document) for several thousand consumer product descriptions, and that the semantic vector representations could be used to find analogies between products, even when they were from different domains (e.g., a "digital collar" that alerts a pet owner when the pet is too far away, and a bluetooth wristband that alerts a parent if the child wearing it is too far away), and at a higher rate than traditional information retrieval approaches (e.g., TF-IDF, GloVe word embeddings [35]).

We are interested in extending Hope et al's [22] approach to help researchers find analogies from the research literature for their own work. For example, could researchers use this approach to find *mechanisms* (e.g., HEATING and gradually COOLING a substance) from other research papers with similar *purposes* (e.g., MINIMIZING imperfections in a substance) that might inspire new ideas for a target *purpose* (e.g., INCREASE and gradually DECREASE the degree of stochasticity in the optimization process to MINIMIZE error in the model, which is the essence of simulated annealing [26])? Or find other papers with the same *purpose* and *mechanism* (e.g., for competitive analysis), even if those papers are from different domains?

However, key features of research papers may prevent a straightforward application of Hope et al's [22] approach. First, research papers are often written in complex, domain-specific language not easily understood by laypeople (e.g., "HMM", "regularization", "cross-validation", "human computation"). The complexity of the concepts and language might greatly increase the cost of human annotation, rendering the method unsuitable for mining analogies across large, complex, multidisciplinary corpora of research papers. Research papers also often include multiple purposes that might have hierarchical causal relationships with each other (e.g., "expose model parameters and error patterns" TO "support algorithmic accountability"). Additionally, many research papers aim to *understand* some phenomenon, instead of contribute novel mechanisms to add value for some user: for these papers, the purpose-mechanism schema might not capture the core relational structure of these papers, and thus fail to find useful analogies for these papers.

In this paper, we explore approaches to address these challenges for extracting and using purpose-mechanism representations to discover analogical relationships in complex scientific literature. Our investigation yields the following contributions:

(1) Annotating the purpose and mechanism aspects of research papers is scalable in terms of cost (<1 minute per paper), not critically dependent on annotator expertise, and generalizable across domains. We explore this by deploying the method with expert annotators on papers in their domain (Study 1) and outside their domain (Study 2), and with novice crowd annotators (Study 3).
(2) Purpose-mechanism representations of research papers can be used to discover analogical relationships that are systematically missed by traditional state-of-the-art semantic models (Study 1), including distant analogies found valuable to a domain expert (Study 2).
(3) Extending the annotation scheme to incorporate other aspects of research papers (e.g., their higher-level problem "background", and the key *findings* of the paper) yields measurable gains in analogy-finding performance (Study 1).

These results suggest that Hope et al's [22] mixed-initiative system (with key modifications from our investigation) is a promising approach for finding analogies amongst research papers. We call this modifed mixed-initiative system **SOLVENT**, to capture its core intution: using a mixture of crowdsourcing and machine learning to "dissolve" research papers into their constituent soft relational schemas (e.g., purpose-mechanism schemas), which can then be used to find analogical combinations of research papers that yield novel discoveries and innovations. We offer **SOLVENT** as a novel path towards computationally supporting knowledge sharing in research communities.

2 RELATED WORK

2.1 Citation Recommendation

The problem of finding analogous research papers can be thought of as a special case of the problem of citation recommendation: given some target manuscript that one is writing, find other papers that are suitable citations [20, 42]. Our work here on finding suitable semantic representations of documents is similar to content-based approaches that leverage semantic representations of documents to search for recommendations[5], and complementary to graph-based approaches (e.g., based on properties of citation/co-authorship networks [29, 37]) to this problem.

2.2 Computational Analogy

Computational analogy is a well-studied problem in cognitive science and artificial intelligence. A major branch of research has devised algorithms for reasoning over rich, relational representations, such as Gentner and colleagues' structure-mapping engine [14, 15], Hummel and Holyoak's LISA analogy engine [23], and Vattam and colleagues' [44] Design Analogy to Nature Engine

(DANE). These systems can achieve impressive, human-like performance on analogy tasks, but require very detailed and rich relational representations: for example, to reason about the simulated annealing analogy, the structure-mapping engine [14, 15] would require a representation of the concept of annealing that looks something like "CAUSE(AND(HEAT(metalworker, metal), CONTROL(metalworker, COOL(metalworker, metal))), MINIMIZE(metalworker, IMPERFECTIONS_IN(metal))". These representations are very costly to obtain: for example, Vattam and colleagues [44] estimated that converting a single (complex) biological system into a formal representation requires between forty and one hundred person-hours of work; further, automatic extraction of relational representations across domains remains a difficult open problem [19].

On the other hand, many computational approaches have been developed that ignore relational structure, opting instead to learn semantic representations from the distributional statistics of words across documents in a corpus—e.g., word embedding models like Word2Vec [32], vector-space models like Latent Semantic Indexing [13], and probabilistic topic modeling approaches like Latent Dirichlet Allocation [6]–scale well to very large datasets, and perform well at matching based on overall/surface similarity, but struggle to detect analogies between documents when surface similarity is low.

Our work extends a new thread of research [22] that explores how to extract and exploit "soft" relational structure: not going all the way to fully specified relational representations, but still attempting to model some kind of relational structure (e.g., semantic representations of the overall purpose and mechanism of a product, which can be used to define a "soft" schema for the product).

2.3 Crowd-Powered and Mixed-Initiative Knowledge Modeling

Many researchers have explored the general problem of applying crowdsourcing techniques to *synthesize* a knowledge model of some domain, from taxonomies of items within a domain [10, 46], to summarizing answers from Web sources for a given question [8, 18], to planning conference talk schedules based in part on topical similarity [1, 9], to organizing ideas for a given problem by topical similarity during collaborative idea generation [40]. Many of these efforts are mixed-initiative systems, where crowds either generate semantic data to support the creation of a computational semantic model [40], or use computational methods—such as information-theoretic methods—to support more efficient crowdsourcing strategies [7, 46], or effectively aggregate crowds' semantic judgments [10]. These applications tend to be about general topicality or similarity of documents/items within some domain, rather than about *analogical* relationships between documents/items that might be from different domains.

Closer to the problem of supporting analogy, other researchers have explored using crowdsourcing to perform or assist with extracting *relations* [30]. Most work in this area aims to extract single commonsense facts (e.g., *"Portland IS_IN Oregon"*, *Obama 'WORKED_AT' the White House*, similar to open information extraction [3], instead of *systems* of relations (e.g., *minimize* imperfections in a metal BY *controlled heating* and *cooling* of the metal): thus, while these relations can be quite useful for improving artificial commonsense reasoning, they are not as useful for reasoning about analogies between scientific papers, especially across multiple disciplines. However, some recent work has examined how to use crowdsourcing to assist with extraction of more specialized relations, for example mechanical engineering-related functions and mechanisms in biological texts [2], or experiment details for biomedical research (e.g., population studied, treatment given) [43].

We are inspired by these efforts (particularly the approach of using mixed-initiative systems), but differ in that we aim to crowdsource knowledge models for complex corpora (like scientific papers) across many different knowledge domains. Additionally, instead of attempting to model specific relations and entities in detail, we aim to extract general semantic categories (e.g., purpose/mechanism) that are relevant for analogical reasoning.

2.4 Ontologies of scientific and scholarly discourse

The general semantic categories we aim to crowdsource are closely related to work on ontologies of scholarly and scientific *discourse* [11, 12], with an especially close relation to efforts like the Core Information about Scientific Papers (CISP) ontology [28], which breaks papers out into "Goal of investigation", "Motivation", "Object of investigation", "Research method, "Experiment", "Observation, "Result", and "Conclusion". While our goal differs from this work in aiming to extract elements that support *analogy* in particular (vs. construct a complete/useful model of scientific discourse), we extend this body of work by systematically exploring how these discourse-like ontology elements might be efficiently crowdsourced (vs. relying entirely on automation or expert researcher annotations [27]), and what particular combinations of these elements might be useful for supporting analogy mining.

3 ADAPTING THE METHOD FOR RESEARCH PAPERS

Applying the purpose-mechanism annotation analogical search method [22] to the domain of research papers may seem straightforward at first glance: most research projects have research goals (e.g., research problems, research questions), and methods for achieving those goals. However, there are several major challenges involving in moving from the domain of simple consumer product descriptions to more complex structures and terminology, such as found in scientific literature.

3.0.1 Complex language. First, research papers tend to be significantly more complex than the typical consumer product idea. For example, most people know that a utility belt is for holding/storing things, so words like "store/keep" are purpose words; in contrast, without content/domain knowledge, it may be difficult to discern which terms denote purpose/mechanism in research papers. For example, a hidden Markov model (HMM) is a common mechanism for analyzing sequences of events. However, without specialized domain knowledge it is possible that an annotator might not be able to label it as a mechanism. Fortunately, as we will show, people are able to leverage rich syntactic cues to determine which *portions* (not specific words) denote the purpose/mechanism/etc of the paper, even in the absence of understanding the details of those purposes or mechanisms.

3.0.2 Hierarchy of problems. Research papers also often address multiple distinct purposes that are hierarchically dependent on each other. For example, a research paper might explore how to improve creativity in teams (high-level problem) by improving the ability of team members' to build on each others' ideas (mid-level subproblem), and specifically aim to investigate ways to surface the most diverse and best ideas (low-level subproblem). In most cases, the mechanisms contributed by a given paper are most directly causally related to the lower level subproblems (e.g., crowdsourcing techniques for summarizing ideas); consequently, the purpose-mechanism schema of a particular paper is often best described by its low-level subproblems, along with the mechanisms it contributes for solving those subproblems.

To address this challenge, it may be useful to distinguish between the research *background* (higher-level problems) and the specific problem(s) being addressed in the paper to find more directly analogous matches. The background context can also be "partialed out" (e.g. by ignoring it) to enable discovery of analogies between different areas/fields, e.g., matching papers that are about crowdsourcing techniques for content analysis in qualitative analysis vs. online classrooms vs. brainstorming teams.

3.0.3 Mechanisms vs. findings. Finally, many research papers do not fit the purpose-mechanism model exactly. Specifically, many papers cannot really be understood in terms of answering "how" questions (e.g., how might we improve work quality for novice crowd workers); instead, their

primary goal is to *understand* some phenomenon (e.g., when and why do people disclose personal information on social media?).

For these "understanding-oriented" papers, the purpose remains informative (what question they want to understand), but the mechanism (how they found the answer) may no longer capture the core contribution of the paper towards that purpose; instead, the *findings* of the paper are the "answer" to the purpose of the paper. For example, a research paper whose *purpose* is to investigate people's motivations for disclosing personal information on social media may have (according to our schema) particular research methods (e.g., interviews with users, content analysis) as *mechanisms*, but the answer to the purpose might be the *findings* it reports (e.g., users avoid disclosing if they want to manage their online identity across multiple social contexts). Thus, the relevant schema to match on might be a purpose-findings schema, or even just a findings schema, as opposed to a purpose-mechanism one.

3.1 Modified annotation scheme

To address these challenges, we adapted Hope et al's [22] purpose-mechanism annotation scheme to incorporate two new elements (background and findings). This yields the following modified annotation scheme with 4 elements:

(1) **Background**: What is the intellectual context of this work? What other (higher-level) goals/questions can be furthered by this work? How might this help other research(ers)?
(2) **Purpose**: What specific thing(s) do the paper's authors want to do or know?
(3) **Mechanism**: How did the paper's authors do it or find out?
(4) **Findings**: Did it work? What did the paper's authors find out?

Table 1 (top) shows examples of these annotations for a system-oriented vs. a more "understanding-oriented" paper. This modified annotation scheme defines the target aspects that SOLVENT transforms into soft relational schemas, which are used for analogical matching.

3.2 Creating semantic vector representations from the annotations

Each of the four annotations (background, purpose, mechanism, and findings) yields a set of words. For example, in Table 1, the set of words highlighted in red are annotated as "purpose". We are interested in taking the four sets of annotations $W = \{w_1, w_2, w_3, w_4\}$ (where each w_i is a set of words), and creating representations that capture the respective aspects of the paper. We follow a common approach in natural language processing and information retrieval, and build semantic *vector* representations based on each annotated set of words w_i. These vector representations are SOLVENT's "soft" relational schemas that can be used for analogical matching.

In particular, we represent words with their *word vectors* (e.g., trained with word2vec [33] or GloVe [35]), and then for each set w_i we compute a weighted average of the words in w_i to obtain a representation for aspect i. The average is weighted by the TF-IDF score of each word vector[2]. Note that the TF-IDF score is calculated on an annotation-specific basis: for example, when computing TF-IDF scores for purpose vectors, we use only words that were tagged as purpose when computing term and document frequencies[3]. This is a conceptually similar approach to that taken by Hope et al [22] and helps us tease apart words that are important for denoting the purpose/mechanism/etc. of documents, as opposed to generalized importance. The end result is that each document is represented by 4 annotation-specific semantic vectors (background, purpose, mechanism, and findings), which we denote as $\mathbf{b}_i \in \mathbb{R}^D$, $\mathbf{p}_i \in \mathbb{R}^D$, $\mathbf{m}_i \in \mathbb{R}^D$, and $\mathbf{f}_i \in \mathbb{R}^D$, respectively (where D is the dimensionality of the base word vectors we use to represent the words).

[2]This step is implemented in `get-vectors.py` in the code provided in the supplementary material
[3]This step is implemented in `compute-tf-idfs.py` in the code provided in the supplementary material

IdeaHound: Self-sustainable Idea Generation in Creative Online Communities	**A Comparison of Social, Learning, and Financial Strategies on Crowd Engagement and Output Quality**
One main challenge in large creative online communities is helping their members find inspirational ideas from a large pool of ideas. A high-level approach to address this challenge is to create a synthesis of emerging solution space that can be used to provide participants with creative and diverse inspirational ideas of others. Existing approaches to generate the synthesis of solution space either require community members to engage in tasks that detract from the main activity of generating ideas or depend on external crowd workers to help organize the ideas. We built IDEAHOUND a collaborative idea generation system that demonstrates an alternative "organic" human computation approach, where community members (rather than external crowds) contribute feedback about ideas as a byproduct of an activity that naturally integrates into the ideation process. This feedback in turn helps the community identify diverse inspirational ideas that can prompt community members to generate more high-quality and diverse ideas.	A significant challenge for crowdsourcing has been increasing worker engagement and output quality. We explore the effects of social, learning, and financial strategies, and their combinations, on increasing worker retention across tasks and change in the quality of worker output. Through three experiments, we show that 1) using these strategies together increased workers' engagement and the quality of their work; 2) a social strategy was most effective for increasing engagement; 3) a learning strategy was most effective in improving quality. The findings of this paper provide strategies for harnessing the crowd to perform complex tasks, as well as insight into crowd workers' motivation.

Top 10 nearest words for

Abstract: ideas idea analogies proposals productively substrates fictions discoveries brainstorm schemas	**Abstract:** efficient expressive retention responsive susceptibility differentially enjoyment disengagement alertness
Background: ideas creative innovative inspirational analogies productively fictions brainstorm substrates wishing	**Background:** crowd workers worker shepherding sourcing harnessing intrinsically legitimacy rubrics extrinsic
Purpose: curate interleave scaffold organise devise stimulate formulate ideas consolidating mobilize	**Purpose:** retention differentially worker soundness mediators maximize susceptibility redefining greatly decreasing
Mechanism: idea ideas ideation endeavor grassroots realization productively scaffold generative prospector	**Mechanism:** nine eleven twelve fourteen fieldwork comprehensively separately uncompensated sixteen thirteen
Findings: ideas community cultivate cohesive heterogeneous divergent productively repertoire mobilize socialize	**Findings:** tactic efficient disengagement responsive playful attentiveness partnerships paradoxically expressive

Table 1. Examples of modified annotation scheme applied to a system-oriented paper (top left) and an "understanding"-oriented paper (top right), and top 10 nearest words for all abstract words VS. each annotation vector for each paper (bottom)

To give an intuition of what the vectors mean, Table 1 (bottom) shows the top 10 nearest neighbor words for each annotation type for two example papers, as well as for a vector representing the full abstract (i.e., TF-IDF-weighted average of all non-stopwords in the abstract). The nearest neighbors are quite different across the different annotation types and from the full abstract vector, and also map well to the meaning of those annotation types. For example, while the overall abstract for the first example is about improving creativity in online communities (as seen in the top words for the abstract vector, *{ideas, idea, analogies, proposals...}*), the core **purpose** of enabling efficient synthesis of the ideas generated by the community is well captured by the words *{curate, interleave, scaffold, organise, devise...}*, and the mechanism of "piggy-backing" on ideation is captured by the **mechanism** top words, *{idea, ideas, ideation, endeavor, grassroots...}*. The term "grassroots" in

particular is interesting because it captures the idea of sustainability, which is completely missed in the top words for the abstract and **background** vectors. In contrast, for the second, understanding-oriented paper, the **mechanism** top words are very research-methods oriented, *{nine, eleven, twelve, fourteen, fieldwork...}*.

3.3 Analogical similarity metrics

Using these specialized vectors, SOLVENT can support a number of interesting analogical queries:

(1) PURPOSE+MECHANISM : find other research papers with a similar *purpose* AND *mechanism*, by computing the average of the cosine between their purpose vectors and the cosine betwen their mechanism vectors, i.e., $\langle \cos(\mathbf{p}_1, \mathbf{p}_2), \cos(\mathbf{m}_1, \mathbf{m}_2) \rangle$ for any pair of $paper_1$ and $paper_2$. By ignoring background/implications, we aim to achieve some level of abstraction away from the domain context.
(2) PURPOSE : find other research papers with a similar *purpose*, but not necessarily a similar *mechanism*, with $\cos(\mathbf{m}_1, \mathbf{m}_2)$ This should find alternative mechanisms for solving some target problem.
(3) MECHANISM : find other research papers with a similar *mechanism*, but not necessarily a similar *purpose*, with $\cos(\mathbf{p}_1, \mathbf{p}_2)$. This should find alternative applications for some mechanism.
(4) FINDINGS : find other research papers with similar *findings*, with $\cos(\mathbf{f}_1, \mathbf{f}_2)$. This should help find analogies between "understanding" papers (where the core schema is not necessarily a purpose-mechanism pairing).

4 STUDY 1: SOURCING ANNOTATIONS FROM DOMAIN-EXPERT RESEARCHERS

We begin our investigation of SOLVENT's quality and feasibility by sourcing and using annotations from domain expert researchers. This deployment should provide an upper bound on SOLVENT's quality and cost considerations.

4.1 Dataset

As an initial test, we deployed our modified mixed-initative system on 50 papers published at the CSCW conference[4] from 2013-2017. This modest number of papers is small enough for experts to manually find analogies (with reasonable confidence that we have covered most analogies), yet large enough to produce interpretable quantitative results (N=1,225 possible pairs between papers). Further, the match to our content/domain expertise (social computing) allows us to better define what counts as analogies, and judge the performance of the different metrics accordingly.

4.2 Annotation and Vectors

Two members of the research team served as annotators. Each paper abstract took a median of 1 minute to annotate. In Studies 2 and 3, we explore how the method extends to settings where annotators lack content/domain expertise.

We originally used pre-trained GloVe [35] vectors trained on the Common Crawl dataset [5]. However, baseline performance was very poor. We therefore opted for a doc2vec model [33], with 600 dimensions, trained on a smaller but more focused and relevant corpus (4,402 papers from CHI and CSCW from 2010 to 2017).

4.3 Baselines

We compare the performance of our analogical similarity metrics to two baselines:

[4]http://cscw.acm.org/
[5]https://nlp.stanford.edu/projects/glove/

Match Type	**Title** & *Explanation*
Near analogy: **Background** + **Purpose** + **Mechanism**	**Toward Collaborative Ideation at Scale: Leveraging Ideas from Others to Generate More Creative and Diverse Ideas**: *Use crowdsourcing to identify diverse idea sets to improve creative idea generation*
Far analogy: **Purpose** + **Mechanism**	**Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms**: *Make worker certification self-sustainable by using crowdsourced peer review to certify good workers*
Non-analogy	**Making Decisions From a Distance: The Impact of Technological Mediation on Riskiness and Dehumanization**: *No similarity*

Table 2. Example expert-found analogies for the IdeaHound paper in the CSCW dataset, with notes on each analogy (in italics). One non-analogy also shown for comparison. Analogies are mapped to the following: core schema from the IdeaHound paper: *Make creative idea generation self-sustainable by using crowdsourced peer review to identify diverse idea sets.*

(1) ALL WORDS baseline: create TF-IDF-weighted vectors for each paper from all the words in each paper's abstract (denoted as $\mathbf{a_i}$), and compute $\cos(\mathbf{a_1}, \mathbf{a_2})$, for every pair of $paper_1$ and $paper_2$. This baseline is meant to emulate current content-based recommendation practices, which often operate on some combination of the abstract/full-text.

(2) BKGROUND/PURPOSE+MECHANISM baseline: create a new semantic vector $\mathbf{bp_i}$ that captures both *background* and *purpose* aspects, by concatenating the weighted vectors for w_b (background tokens) and w_p (purpose tokens; each weighted by their respective annotation-specific tf-idf weights), and computing the average of this concatenation. We then look for matches with similar *background*/*purpose* and *mechanism*, i.e., $\langle \cos(\mathbf{bp}_1, \mathbf{bp}_2), \cos(\mathbf{m}_1, \mathbf{m}_2) \rangle$. Comparing our new metrics (especially the PURPOSE+MECHANISM and PURPOSE metrics, which ignore *background*) to this baseline tests whether it is necessary to separate *background* and *purpose*, or whether it is sufficient to simply group all the *purposes* of a given paper as a single annotation type (inspired by the purpose-mechanism method from Hope et al. [22]).

4.4 Performance Measures

To evaluate the performance of the similarity metrics, we use *Precision@K*, defined as the number of known analogies in the dataset found in the top K% of matches. Exploring different levels of K allows us to explore how the performance of a given metric changes depending on how conservative it is (lower K = more conservative, which accepts a higher risk of false negatives in favor of a higher precision of matches returned to the searcher).

These analogies were manually found by a member of the research team through exhaustive examination of the papers in two phases. First, the research team member created affinity maps of the research papers, and identified shared schemas and relations. Second, these schemas were used to find papers that fit the schema (thereby defining analogy pairs). Finally, we used an initial prototype GloVe model (with Common Crawl) to suggest new matches we might have missed. We ended up with 259 analogy pairs (approximately 21% of 1225 total possible pairs across 50 papers; see Table 2 for some example analogies found for the IdeaHound paper from Table 1[6]).

The research team member who found the analogies in the dataset has 8 years of PhD-level expertise in researching social computing and analogy, providing both domain and process (for finding analogies) expertise. The team member considered papers to be analogically related if they share a relational mapping (e.g., similar purposes and mechanisms, or similar backgrounds and

[6]The full set of known analogy pairs is available in the supplementary material

	K = 1%	2%	5%	10%	15%	20%	25%
ALL WORDS baseline	.67 *(.03)*	.67 *(.06)*	.46 *(11)*	.37 *(.18)*	.35 *(.25)*	.31 *(.29)*	.29 *(.34)*
BKGROUND/PURPOSE+MECHANISM	.84 *(.05)*	.84 *(.09)*	.64 *(.16)*	**.50** *(.24)*	.43 *(.30)*	.37 *(.35)*	.33 *(.39)*
PURPOSE+MECHANISM	**.92** *(.05)*	**.92** *(.09)*	**.73** *(.18)*	**.50** *(.24)*	.40 *(.29)*	.36 *(.34)*	.34 *(.40)*
PURPOSE	.50 *(.03)*	.50 *(.05)*	.38 *(.10)*	.38 *(.18)*	.33 *(.23)*	.31 *(.29)*	.30 *(.35)*
MECHANISM	**.92** *(.05)*	.80 *(.08)*	.66 *(.16)*	.49 *(.23)*	**.45***(.32)*	**.39** *(.37)*	**.35** *(.42)*
FINDINGS	.75 *(.04)*	.50 *(.05)*	.37 *(.09)*	.33 *(.16)*	.31 *(.22)*	.29 *(.28)*	.27 *(.32)*

Table 3. Proportion of known analogy pairs found in Study 1 for each metric, varying by K (lower K = more conservative; recall scores in parentheses; pairs sorted by similarity). Best-performing scores at each K are **bold-underlined**. Our BKGROUND/PURPOSE+MECHANISM , PURPOSE+MECHANISM , and MECHANISM metrics consistently find analogy pairs at substantially higher precision than the ALL WORDS baseline at each level of K.

purposes). For example, motivating contributions in citizen science relationally maps to motivating scientists to share their code (e.g., motivating(x to contribute to y)). Papers with similar backgrounds (e.g., vision science) but different purposes and mechanisms would not be considered analogies. This is consistent with the literature on analogical mapping (e.g., [15]). Importantly, two papers with similar background would still be considered analogies if their purposes or mechanisms mapped relationally; these are known in the analogy literature as near analogies.

Note that because these are all social computing papers, many of the analogies are from the same/similar domains/areas; therefore, we expect the ALL WORDS metric to be able to find many of these near analogies. Separating **background** from **purpose** also may not be as important for finding near analogies. For this reason, our performance measure in this dataset is a conservative test of the efficacy of our mixed-initiative system. Yet, social computing is broad enough as a field that it allows for some cross-domain analogies (e.g., crowdsourcing grades for assignments in MOOCs, vs. crowdsourcing discovery of creative ideas in online innovation platforms).

We examine precision for K ∈ (1, 2, 5, 10, 15, 20, 25). For each similarity metric, and for each K, we compute similarities between all 1,225 possible pairs in the dataset, rank based on those similarities, take the top K% of matches, and then compute how many of those matches were known analogies.

4.5 Results

4.5.1 We find unique analogies missed by the ALL WORDS baseline. Table 3 shows the quantitative results of our experiment. Three of our similarity metrics (BKGROUND/PURPOSE+MECHANISM , PURPOSE+MECHANISM and MECHANISM) consistently return a higher proportion of analogies at multiple settings of K, yielding on average gains of 26%, 31%, and 30% in precision compared to ALL WORDS . The advantage of our metrics is especially substantial at lower levels of K: for example, in the top 1% and 2% of matches, our best metric (PURPOSE+MECHANISM) has both a large increase over the ALL WORDS baseline (a 37% increase from .67 to .92), and a high precision value in absolute terms. This is significant because we did not distinguish between near and far analogies, and there were many near analogies in the dataset; these near analogies are likely to dominate at lower levels of K for ALL WORDS (which is tuned largely for surface similarity).

Additionally, there was low overlap between the matches found by our metrics and those found by ALL WORDS : for example, only 42% of the PURPOSE+MECHANISM , and 50% of the MECHANISM matches were shared with the ALL WORDS baseline at K=5. Referring back to the matches in Table 2, the ALL WORDS baseline was only able to find the near analogy ("Towards Collaborative Ideation at

Source: IdeaHound: Self-sustainable Idea Generation in Creative Online Communities	**Analogy**: Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms
Top 10 nearest words for	
Background: ideas creative innovative inspirational analogies productively fictions brainstorm substrates wishing	**Background:** workers decentralized crowd crowds dispersed sourcing equitably collaborators volunteers harness
Purpose: **curate** interleave scaffold **organise** devise stimulate formulate ideas consolidating mobilize	**Purpose**: **reputation scores score ratings** accountability legitimacy fairness responsiveness **scoring relevancy**
Mechanism: idea ideas ideation endeavor **grassroots** realization productively scaffold generative prospector	**Mechanism**: crowd workers crowds guilds worker **peer sourcing** instructors equitably freelancing

Table 4. Illustrative analogy (right) found for the IdeaHound paper (left) in the CSCW dataset by our PURPOSE+MECHANISM and PURPOSE metrics, but missed by the ALL WORDS and BKGROUND/PURPOSE+MECHANISM baseline metrics. **Bold-underlined** words denote conceptual overlap across matching vectors. Note how different the background vectors are, while there is overlap in concepts for both purpose and mechanism vectors. Both papers leverage community mechanisms [grassroots, peersourcing] for the purpose of curating things [ideas, workers], despite key differences in background (creativity/ideation vs. crowd labor markets). This example illustrates the value of separating higher-level (background) problems from lower-level (purpose) problems during annotation.

Scale"), while the other metrics were able to find both the near and far analogy ("Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms").

4.5.2 Modifications to the annotation scheme are helpful. While BKGROUND/PURPOSE+MECHANISM finds more analogies than ALL WORDS across all levels of K, it is consistently outperformed by PURPOSE+MECHANISM (which separates background and purpose) across all levels of K.

Inspection of the matches suggests that separating background and purpose allows us to find analogies between papers that have different higher-level goals (and might be thought of as in different domains). Table 4 shows one example of this, where PURPOSE+MECHANISM found an analogy between a paper on devising scalable methods to organise many ideas from an online creative community and distribute them to inspire community members ("IdeaHound: Self-Sustainable Idea Generation in Creative Online Communities"), and a paper on using peer ratings to rank workers by quality in online labor markets ("Crowd Guilds: Worker-led Reputation and Feedback on Crowdsourcing Platforms"). This match was missed by both ALL WORDS and BKGROUND/PURPOSE+MECHANISM , in part because the background vectors are quite different from each other (one heavily emphasizing creativity/ideation, and the other heavily emphasizing crowd work).

4.5.3 Findings annotations help find near analogies. While the FINDINGS metric does return significantly more matches in the top 1% of pairs, it does not outperform the ALL WORDS baseline at K>=2. This is partially explained by the high amount of overlap between the matches found by FINDINGS and those found by the ALL WORDS baseline metric (64% are shared, in contrast to the low levels of overlap for the PURPOSE+MECHANISM and MECHANISM metrics). We will return to this point in the discussion.

5 STUDY 2: USING SOLVENT TO FIND ANALOGIES WITH REAL-WORLD VALUE

Study 1 demonstrated that applying **SOLVENT** to a corpus of social computing papers enabled us to retrieve known analogical relationships, as defined by a domain expert. In Study 2, we explore

whether SOLVENT can provide real-world value for researchers actually looking for analogical inspirations for their work.

5.1 Scenario

We evaluated our approach with a local mechanical engineering research group at a highly research active, private research university in the Midwestern United States. The group is working on cutting-edge interdisciplinary work at the intersection of bioengineering and mechanical engineering: specifically, they are exploring how to create interesting new 2D and 3D structures at many different scales by stretching/folding polymers.

The research group has been struggling to find examples of this work to build on and compare against: it is distributed across such diverse domains as mechanical engineering, civil engineering, aerospace engineering, materials science, mathematics, and design, among others, many of which are outside of the specific domain expertise of the research group.

Some examples of their analogy information needs include:

- *Competitive analysis*: find work from other research groups that are attacking the same problems with similar techniques
- *Inspiration*: find relevant research that can suggest new properties of polymers to leverage, or new techniques for stretching/folding or exploring 2D/3D structure designs
- *Application*: find novel application domains that could benefit from the unique advantages of their fabrication method (e.g., non-invasively constructing temporary medical implants *in situ*)

The team has spent the last year or so conducting literature searches using various standard approaches (e.g., Google Scholar, library databases, citations in relevant [review] papers, conversations with colleagues), and have been frustrated by how slow and error-prone the process has been. They are also concerned that they are missing things. The difficulty stems in part from the relative newness of the field, and the degree of fragmentation of knowledge across many different disciplines.

In this evaluation, we explore whether SOLVENT can help them find analogies they were not able to find through keyword/database and citation tree search.

5.2 Dataset

Based on conversations with the group, we identified three out-of-domain sources of research papers to mine from: materials science, civil engineering, and aerospace engineering. The papers were sampled from the 1000 most highly cited papers of 2016-2017 (indexed on Web of Science) for each of the 3 domains. Two members of our research team (neither of whom have any training or formal knowledge in any of the domains involved in this scenario) annotated 90 papers sampled from this larger corpus.

As before, rather than use a semantic model trained on a larger but more generalized corpus (e.g., Common Crawl), we used a model trained on a smaller but more relevant corpus of documents to create aggregate vectors. Specifically, we used word vectors from a continuous-bag-of-words (CBOW) word2vec model [32] trained on 3,000 papers in the dataset. We then used the same method as before to create annotation-specific semantic vectors for each of the 90 annotated papers.

5.3 Evaluation

We used an abstract for one of the research group's recent manuscripts as a query document. We first highlighted and constructed semantic vectors for the abstract, and then sampled the top 20 matches for the abstract from each of our similarity metrics from before. This corresponds to a

setting of K at approximately 20% (top 20 most similar out of 90 possible matches); we chose this less conservative setting of K to reflect our prior expectation that the number of cross-domain matches that would be useful to the team is likely to be relatively few. We collapsed duplicates and blinded them by removing information about which metric produced the match, resulting in a set of 53 unique possible matches.

We showed these matches to a member of their research group (the lead PhD student). We asked the researcher look through the matches as s/he might look through a Google Search result list, and evaluate 1) whether each match is in fact useful for their literature review, 2) why (not), and 3) whether they had previously encountered that match.

5.4 Results

5.4.1 Our metrics find more analogies than ALL WORDS baseline. Overall, the researcher identified 7 out of the 53 matches as relevant and new. This reflects the extremely challenging setting of finding useful and novel analogies among 90 randomly sampled papers from outside the core domain of the work. 5 out of those 7 analogies were found by one of/both the PURPOSE+MECHANISM and MECHANISM similarity metrics (an aggregate precision of approximately .25; comparable to precision at K=.2 in Study 1), while 2 were found by the ALL WORDS metric (precision = .1).

5.4.2 Our metrics find different analogies from ALL WORDS baseline. In addition to this numerical difference, there was also a qualitative difference in the types of matches returned. To illustrate, the most relevant match found by the ALL WORDS metric was almost a direct replication of their work (manipulating the structure of a membrane in a controlled way by applying tension to the membrane) from an aerospace engineering paper. The researcher was uncertain if she had seen this paper before, and found it useful for *competitive analysis*.

In contrast, the most relevant match found by the PURPOSE+MECHANISM metric was a paper on using multi-agent systems to generate a variety of geometric structures. The MECHANISM metric also found an interesting potential analogy of a civil engineering paper analyzing web crippling phenomena in steel beams under tension. Both of these matches were judged to be both potentially useful (the first for *inspiring* alternative approaches to systematically explore 2D/3D structures by applying different levels and types of tension, the second for potentially *inspiring* new mathematical models to analyze their polymers, which could reveal new properties to leverage for their designs) *and* novel (neither had been seen by the researcher).

6 STUDY 3: SCALING UP SOLVENT THROUGH CROWDSOURCING

Thus far we have shown that **SOLVENT** increases our ability to find known analogies, as well as analogies that can provide real-world value to a domain expert. However, in both Study 1 and 2, our annotations were sourced from expert researchers, who make up a highly limited resource pool. In this final section, we explore how we might scale up **SOLVENT** by crowdsourcing annotations from a wider range of non-experts who might provide a much larger pool of possible human judgments.

Study 2 (where researchers annotated papers outside of their domain/discipline) suggests that annotation does not critically depend on domain knowledge; still, non-experts with little to no experience writing or reading research papers as a "genre" may produce annotations that are too noisy to support effective analogy-mining (or require expensive quality control mechanisms that would make the cost of crowdsourcing prohibitively high). Therefore, in Study 3, we explore whether crowdsourced annotations for papers in the set of 50 CSCW papers can replicate or approximate the analogy-finding gains shown in Study 1 with researcher annotations. We crowdsource annotations from workers on Upwork (where we recruited workers with general writing expertise but no

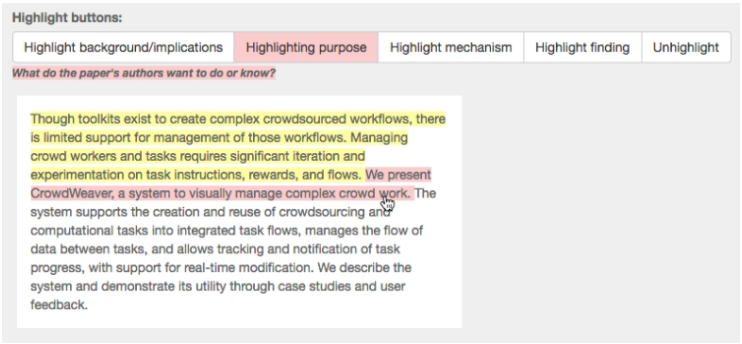

Fig. 1. A screenshot of the crowd-facing interface for annotating the abstracts. Workers annotate aspects by first selecting the annotation type (an in-line description is provided as a reminder) and then using an intuitive click-and-drag highlighting interaction to annotate one or more words.

domain or research experience) and workers on Amazon Mechanical Turk (who represent a larger population of workers with little or no expertise requirements).

6.1 Crowdsourcing setup

Workers used an intuitive drag-and-highlight interaction to annotate each abstract (see Figure 1). We included a training step before the main task: 2 gold standard examples to illustrate the annotations in context, and a single training example (where workers would first annotate and then compare with the gold standard).

6.1.1 Deployment on Upwork. We recruited two workers with general copywriting expertise from Upwork. The first worker held a bachelor's degree in English Literature, and a master's degree in business administration, had more than nine years experience writing and editing academic, business, literary and technical documents, and was paid at a rate of $30/hr. The second worker held a master's degree in English Literature and a certificate in Publishing, had experience with editing book-length projects for major publishers and creating research dossiers for news publications, and was paid at a rate of $20/hr. Annotation times were similar to the MTurk and Study 1 deployments, with a median completion time per document of 1 minute. Each worker provided annotations for half (25) of the papers in the dataset. Thus, each paper's annotations was based on one worker, similar to the deployment with domain-expert researchers in Study 1.

6.1.2 Deployment on Amazon Mechanical Turk. We screened for workers with at least 95% approval rate for at least 5,000 tasks. The overall task (including login, training, and the actual task) took a median of 4.0 minutes; actual time annotating the main document took a median of 1.3 minutes. We paid workers $0.70 for each task completion, for an effective average hourly rate of $10/hr. We obtained annotations from an average of 3 workers for each document. We aggregated annotations across workers for each word by majority vote (weighted by each worker's performance on the training example).

6.2 Results

6.2.1 Crowd annotations had substantial agreement with researcher annotations. Table 5 shows the accuracies (i.e., agreement with researcher annotations) by annotation type. Overall, Upwork

	Overall	Bkgd	Find	Mech	Purp
Upwork	**0.78**	0.75	0.94	0.72	0.68
MTurk	**0.59**	0.66	0.71	0.53	0.42

Table 5. Crowd annotation agreement with researcher annotations

	K = 1%	2%	5%	10%	15%	20%	25%
ALL WORDS	.67 *(.03)*	.67 *(.06)*	.46 *(11)*	.37 *(.18)*	.35 *(.25)*	.31 *(.29)*	.29 *(.34)*
BKGROUND/PURPOSE+MECHANISM							
Upwork	.75 *(.04)*	.67 *(.06)*	**.64** *(.15)*	**.43** *(.20)*	**.39** *(.28)*	**.33** *(.31)*	**.31** *(.37)*
MTurk	.75 *(.04)*	.63 *(.06)*	.57 *(.14)*	**.43** *(.20)*	.34 *(.24)*	.32 *(.30)*	.29 *(.34)*
PURPOSE+MECHANISM							
Upwork	**.83** *(.04)*	**.75** *(.07)*	.54 *(.13)*	.42 *(.20)*	.34 *(.24)*	**.33** *(.31)*	.30 *(.35)*
MTurk	**.83** *(.04)*	.71 *(.07)*	.52 *(.13)*	.42 *(.20)*	.37 *(.26)*	.32 *(.30)*	**.31** *(.36)*
MECHANISM							
Upwork	.75 *(.04)*	.58 *(.06)*	.51 *(.12)*	.39 *(.19)*	.33 *(.23)*	.31 *(.30)*	.30 *(.36)*
MTurk	.75 *(.04)*	.54 *(.05)*	.44 *(.11)*	.34 *(.16)*	.32 *(.23)*	.29 *(.27)*	.28 *(.33)*

Table 6. Proportion of known analogies found at each level of K for each similarity metric, with annotations by Upwork and MTurk crowd workers (recall in parentheses; best-performing scores **bold-underlined**. Our BKGROUND/PURPOSE+MECHANISM and PURPOSE+MECHANISM metrics continue to outperform the ALL WORDS baseline (with substantial reductions from researcher-based annotations), while the MECHANISM metric's advantage is almost entirely eliminated for both Upwork and MTurk annotations.

workers' annotations matched the researcher annotations 78% of the time, and MTurk workers' annotations matched 59% of the time. There was considerable variation across papers, with accuracies as high as 96% agreement for some papers, and as low as 4% for others.

6.2.2 Crowds struggled with purpose and mechanism annotations. In general, both groups of workers struggled most with purpose and mechanism annotations (68% and 72% for the Upwork workers, and 42% and 53% agreement for the MTurk workers). First, workers often confused background and purpose. The most frequent error for background annotations was incorrectly annotating it as purpose (13% of background annotations for Upwork; 21% for MTurk). Purpose was also frequently confused for background (10% of purpose annotations for Upwork; 16% for MTurk).

Mechanism was also frequently confused for purpose and findings: the most frequent error for purpose was incorrectly annotating it as mechanism (17% of purpose annotations for Upwork; 28% for MTurk), while mechanism was confused for purpose and findings quite frequently as well (8% and 9% of mechanism annotations for Upwork; 10% and 31% for MTurk). One possible explanation is that researchers frequently mix multiple aspects in their sentences: for example, they often described the purpose and mechanism of their paper in the same sentence, or described the mechanism in more detail by claiming that it worked (findings). We return to this point in the discussion to consider how our scheme might be modified to account for variations in writing style.

6.2.3 Crowd annotations still improve analogy-mining. As Table 6 shows , despite the modest levels of agreement, specialized vectors constructed from crowd workers' aggregated annotations were still able to support BKGROUND/PURPOSE+MECHANISM and PURPOSE+MECHANISM metrics that

outperformed the ALL WORDS baseline, albeit with substantial reductions in the size of the advantage. In particular, the PURPOSE+MECHANISM metric outperforms ALL WORDS throughout the range of K, for both Upwork and MTurk deployments.

Note that the advantage of the specialized metrics is reduced the least for BKGROUND/PURPOSE+MECHANISM (retaining substantial advantages over ALL WORDS , and even outperforming PURPOSE+MECHANISM , past K=2), while the advantage of the MECHANISM metric disappears past K=1. The generally low accuracy for purpose and mechanism annotations might explain these patterns, since the PURPOSE+MECHANISM and MECHANISM metrics depend critically on these annotations.

7 DISCUSSION

7.1 Summary and Implications of Findings

Across three studies, we show that applying **SOLVENT** to research papers yields a substantial improvement (approximately 25-30% overall improvement in Study 1) over a state-of-the-art information retrieval (IR) technique (TF-IDF-weighted average semantic vectors) for finding analogies between those research papers. The advantage holds for near analogies (where standard IR is expected to perform well), and across domains (seen especially in Study 2 with the mechanical engineering research group). We further show that these annotations do not require extensive content/domain expertise to be useful: researchers can produce useful annotations for papers outside their discipline (Study 2), and naive crowd workers can produce useful annotations for research papers (Study 3; albeit with noticable reductions in effectiveness compared to researcher-sourced annotations). Across these deployments, we find that the time cost of annotation is approximately a minute per paper.

Our modified annotation scheme (and the usefulness of partialing out background, or leveraging findings) points to **SOLVENT**'s core insight (which extends Hope et al's [22] approach): "dissolving" documents into relational elements (e.g., purpose, mechanism, findings) that have in-built causal relations with each other enables us to create **soft relational schemas** that are useful for analogical matching. This insight might prove useful in other domains than research papers: for example, annotating the *precedents*, *facts*, and *decisions* in legal cases to support case law reasoning, annotating student *learning goals* and associated *exercises* and/or *examples* in lesson plans to support innovation in teaching, or annotating *plot trajectory* and associated *tropes* in writing. While we suspect that **SOLVENT** might be useful in many of these domains, we think that domains where these dimensions are hidden/obscured by surrounding texts that serve many other functions (e.g., preambles in legal opinions) might especially benefit from **SOLVENT**'s soft schema extraction techniques.

This insight opens up a design space that could help resolve the cost-accuracy tradeoff for computational analogy: rather than spending prohibitively high amounts of resources to specify rich relational structures manually (thereby maximizing accuracy for analogical matching), or completely automating knowledge modeling (but ignoring structure, thereby losing accuracy for analogical matching), we can explore the middle ground of extracting soft relational schemas in a cost-effective manner. This design space mirrors current Semantic Web efforts that seek to relax formality requirements for shared ontologies, and explore how machine- and crowdsourced semantics might complement or even replace more formal and precise (but infeasible to obtain for the broader Web) ontologies [4].

7.2 Limitations and Future Work

7.2.1 Extending to larger datasets. While our results show significant promise, we acknowledge that our datasets are relatively small. This is partially mitigated by the diversity of domains represented: our data includes papers from diverse domains and contribution types/framings (e.g.,

civil engineering papers on numerical analysis of beam strength used a different structure than empirical CHI paper abstracts). However, we do aim to test this approach across larger and more diverse corpora. Part of the bottleneck is having a good set of gold standard matches that we can use to evaluate our approach: conservatively, many datasets might contain approximately 5x as many analogy pairs as there are papers (similar to the 259 known analogy pairs in our set of 50 papers in Study 1); finding analogies by hand remains a costly process, and having relatively complete coverage of known analogies is critical for meaningful precision and recall performance measures. However, alternative evaluations (e.g., experiments with searchers, real-world prototype deployments) are possible.

Extending the method to larger datasets may require increases in the scalability of the method. However, our findings with the crowdsourcing study suggest that obtaining annotations for more realistic-sized datasets may not be cost-prohibitive: for example, our deployments suggest that annotating a corpus of 10,000 papers would require on the order of 10–30,000 person-minutes (or 166-500 person-hours) of work. This suggests that such a corpus could be annotated for as little as under $4,000 (assuming a single trained full-time employee working at $20/hr for 1-2 months, similar to one of our Upwork workers), quite possibly a significant bargain given the high potential upside of transformative cross-disciplinary breakthroughs and reductions in redundant work and dead ends. This work could also be crowdsourced for a comparable cost (and significantly quicker turnaround time), although more sophisticated aggregation and quality control mechanisms than majority vote (see, e.g., [39]), or worker-centric screening/training strategies [34] might be necessary for crowdsourcing efforts similar to our MTurk deployment. Further, with a few thousand papers, (semi) supervised machine learning models like recurrent neural networks (e.g., as in Hope et al [22]) or conditional random-field models might be trained to replace or augment human judgments.

As noted earlier, we are also intrigued by the possibility of more "community-sourced" models of crowdsourcing, where community members contribute semantic annotations as a seamless part of the primary task they are motivated to do (e.g., bookmarking important steps in a how-to video [25], or arranging ideas on a virtual whiteboard in the natural course of brainstorming [40]). In this spirit, could we design a future where the many hours that scientists spend reading, annotating, organizing, and summarizing research papers (e.g., for literature reviews, or peer review) could contribute not just to their primary task of doing research, but also towards supporting a computational infrastructure for knowledge sharing? These approaches could pave the way towards more *self-sustainable* computational infrastructures for knowledge sharing. Note also that in some fields (e.g., human factors, biomedical research), structured abstracts are commonly produced by researchers, suggesting that the publication process could be modified slightly to *author-source* these annotations at a tolerable/minimal cost to authors, particularly if the value of these annotations was made apparent to the authors.

7.2.2 Ensuring unbiased coverage. Recall that we modified the annotation scheme (e.g., adding a *findings* annotation) to reflect the different epistemological goals that exist among research papers (e.g., understanding vs. system-building). We had mixed success with the *findings* annotations: at the highest level of matching (top 1% of matches), the metric outperformed the ALL WORDS baseline, but not at higher levels of K. One possible explanation is that our modest-sized dataset didn't contain enough "same-findings" analogies to have high precision at higher levels of K. Another possible explanation is that our core insight of getting a schema "for free" by identifying separate sets of annotations that are joined by a causal relation glosses over the relational structure *within* an annotation type - in the case of understanding-oriented papers, the relevant schema may consist of relations between objects as hypothesized in the *purpose* statement, or in the *findings* (e.g., people

decide when to self-disclose on social media *depending on* how the disclosure might reflect on their online identity).

Relatedly, our conversations with the mechanical engineering researchers revealed that the *properties* of the objects (e.g., the particular ways that polymers respond to physical loads) are a key aspect of identifying *relevant* analogies to work in other domains. Unfortunately, these properties are often left implicit in the text. This challenge could be addressed by augmenting the vector representations of the documents with key properties (e.g., from knowledge bases like ConceptNet [7] and CYC[8]). How to select which among the numerous possible properties of an entity might be useful for analogical matching would be an interesting and challenging problem for future work; some recent work is already beginning to explore this for consumer product descriptions [17]

More generally, we want to ensure that relying on this annotation schema does not cause us to systematically miss or misrepresent the contributions of papers written in different genres (e.g., that don't emphasize explaining mechanisms, or emphasize reporting findings instead of clearly stating the problem being addressed), or with different kinds of contributions (e.g., conceptual/theory or review papers), or levels of writing/communication quality. We did notice some limitations in our deployments, particularly in trying to apply a purpose/mechanism to review/survey papers, when mechanisms were dropped in favor of motivating the problem in shorter abstracts (possibly with strict word limits), or when the research problem was not clearly/explicitly stated; anecdotally, we noticed that this last writing style (missing problem statements, but rich descriptions of findings and methods) seemed to be especially common in our materials science and engineering papers; this might partially explain the lower performance of our metrics in Study 2 (in addition to the difficulty of explicitly searching in randomly sampled papers from other domains). Future work that builds on this should explore how we might expand the annotation schemes to fit different genres/contributions (while retaining the central insight of identifying soft schemas), or supplement the knowledge models constructed by our methods with alternative methods that are less dependent on the way a paper is written.

7.2.3 Usefulness in Real-World Settings. Finally, while our metrics outperformed our instantiation of a traditional state-of-the-art content-based information retrieval approach (i.e., leveraging word embedding representations of the whole abstracts), our absolute precision was high only for very conservative (low) levels of K (e.g., recommending only the top 1% most similar pairs as analogies). In smaller search spaces or for niche topics with fewer "true" matches, potential users of our approach might have to work harder to find good matches; in many cases, however, we believe that even returning the top 1% most similar matches might return more potential matches than the average user would want to sift through.

Our approach also differs significantly from other production systems in that it ignores other important signals available from the citation graph (e.g., quality, relatedness). In practice, we think it would be useful to combine our content-based approach with graph-based approaches [42], e.g., re-ranking with graph-based signals after retrieving initial candidates via our approach, or re-ranking using our approach after retrieving initial candidates via graph-based approaches.

8 CONCLUSION

In this paper, we introduced **SOLVENT**, a mixed-initiative system for re-representing and finding analogies between research papers across different domains, in which humans produce lightweight annotations of key aspects of research paper abstracts, and a computational model constructs semantic vectors from the annotations and uses them to finds analogies between the papers. We

[7] http://conceptnet.io/
[8] http://www.opencyc.org/

show that **SOLVENT** holds promise for efficiently finding analogies between research papers across domains, opening up novel pathways towards computationally augmenting knowledge sharing for scientific progress.

9 ACKNOWLEDGEMENTS

This work was funded by NSF grants CHS-1526665 and CHS-1816242, ISF grant 1764/15, the HUJI Cyber Security Research Center in conjunction with the Israel National Cyber Bureau in the Prime Minister's Office, Carnegie Mellon University's Web2020 Initiative, and Bosch Research Institute. We thank Yla Tausczik and Ping Wang for comments on earlier drafts of this manuscript.

REFERENCES

[1] Paul AndrÃĭ, Haoqi Zhang, Juho Kim, Lydia Chilton, Steven P. Dow, and Robert C. Miller. 2013. Community clustering: Leveraging an academic crowd to form coherent conference sessions. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[2] Ryan Arlitt, Friederich Berthelsdorf, Sebastian Immel, and Robert B. Stone. 2014. The Biology Phenomenon Categorizer: A Human Computation Framework in Support of Biologically Inspired Design. *Journal of Mechanical Design* (2014). https://doi.org/10.1115/1.4028348

[3] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open Information Extraction from the Web.. In *IJCAI*, Vol. 7. 2670–2676.

[4] Abraham Bernstein, James Hendler, and Natalya Noy. 2016. A New Look at the Semantic Web. *Commun. ACM* 59, 9 (Aug. 2016), 35–37. https://doi.org/10.1145/2890489

[5] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. *arXiv preprint arXiv:1802.08301* (2018).

[6] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* (2003), 993–1022.

[7] Jonathan Bragg and Daniel S. Weld. 2013. Crowdsourcing Multi-Label Classification for Taxonomy Creation. In *First AAAI Conference on Human Computation and Crowdsourcing*.

[8] Joseph C. Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with crowds and computation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM.

[9] Lydia B. Chilton, Juho Kim, Paul AndrÃĭ, Felicia Cordeiro, James A. Landay, Daniel S. Weld, Steven P. Dow, Robert C. Miller, and Haoqi Zhang. 2014. Frenzy: Collaborative Data Organization for Creating Conference Sessions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1255–1264. https://doi.org/10.1145/2556288.2557375

[10] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1999–2008. https://doi.org/10.1145/2470654.2466265

[11] Paolo Ciccarese, Elizabeth Wu, Gwen Wong, Marco Ocana, June Kinoshita, Alan Ruttenberg, and Tim Clark. 2008. The SWAN biomedical discourse ontology. *Journal of Biomedical Informatics* 41, 5 (Oct. 2008), 739–751. https://doi.org/10.1016/j.jbi.2008.04.010

[12] Tim Clark, Paolo N. Ciccarese, and Carole A. Goble. 2014. Micropublications: a semantic model for claims, evidence, arguments and annotations in biomedical communications. *Journal of Biomedical Semantics* 5 (July 2014), 28. https://doi.org/10.1186/2041-1480-5-28

[13] Scott Deerwester, Susan T. Dumais, Geroge W. Furnas, and Thomas K. Landauer. 1990. Indexing by Latent Semantic Analysis. *JASIST* 41, 6 (1990), 1990.

[14] Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence* 41, 1 (1989), 1–63.

[15] Dedre Gentner. 1983. Structure-Mapping: A Theoretical Framework for Analogy*. *Cognitive science* 7, 2 (1983), 155–170.

[16] M. L. Gick and K. J. Holyoak. 1983. Schema induction and analogical transfer. *Cognitive Psychology* 15, 1 (1983), 1–38.

[17] Karni Gilon, Joel Chan, Felicia Y Ng, Hila Lifshitz Assaf, Aniket Kittur, and Dafna Shahaf. 2018. Analogy Mining for Specific Design Needs. In *Proceedings of the 2018 ACM SIGCHI Conference on Human Factors in Computing*.

[18] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2258–2270. https://doi.org/10.1145/2858036.2858364

[19] Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet Semantic Role Labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Vol. 1. 471–482.

[20] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 421–430. https://doi.org/10.1145/1772690.1772734

[21] K. J. Holyoak and P. Thagard. 1996. The analogical scientist. In *Mental Leaps: Analogy in Creative Thought*, K. J. Holyoak and P. Thagard (Eds.). Cambridge, MA, 185–209.

[22] Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating Innovation Through Analogy Mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 235–243.

[23] John E Hummel and Keith J Holyoak. 2003. A symbolic-connectionist theory of relational inference and generalization. *Psychological review* 110, 2 (2003), 220.

[24] Benjamin F. Jones. 2009. The Burden of Knowledge and the Death of the Renaissance Man: Is Innovation Getting Harder? *Review of Economic Studies* 76, 1 (2009), 283–317. https://doi.org/10.1111/j.1467-937X.2008.00531.x

[25] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing Step-by-step Information Extraction to Enhance Existing How-to Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 4017–4026. https://doi.org/10.1145/2556288.2556986

[26] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. 1983. Optimization by simulated annealing. *science* 220, 4598 (1983), 671–680.

[27] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics* 28, 7 (April 2012), 991–1000. https://doi.org/10.1093/bioinformatics/bts071

[28] Maria Liakata, Simone Teufel, Advaith Siddharthan, Colin R Batchelor, and others. 2010. Corpora for the Conceptualisation and Zoning of Scientific Papers.. In *LREC*. Citeseer.

[29] Yicong Liang, Qing Li, and Tieyun Qian. 2011. Finding Relevant Papers Based on Citation Relations. In *Web-Age Information Management (Lecture Notes in Computer Science)*. Springer, Berlin, Heidelberg, 403–414. https://doi.org/10.1007/978-3-642-23535-1_35

[30] Angli Liu, Stephen Soderland, Jonathan Bragg, Christopher H Lin, Xiao Ling, and Daniel S Weld. 2016. Effective Crowd Annotation for Relation Extraction.. In *HLT-NAACL*. 897–906.

[31] Salvador E Luria and Max Delbrück. 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28, 6 (1943), 491.

[32] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]* (Jan. 2013). http://arxiv.org/abs/1301.3781 arXiv: 1301.3781.

[33] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3111–3119.

[34] Tanushree Mitra, C.J. Hutto, and Eric Gilbert. 2015. Comparing Person- and Process-centric Strategies for Obtaining Quality Data on Amazon Mechanical Turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1345–1354. https://doi.org/10.1145/2702123.2702553

[35] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12 (2014), 1532–1543.

[36] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[37] Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 2014. ClusCite: effective citation recommendation by information network-based clustering. In *Knowledge Discovery and Data Mining*. 821–830. https://doi.org/10.1145/2623330.2623630

[38] R. Keith Sawyer. 2012. *Explaining creativity: the science of human innovation* (2nd ed.). Oxford University Press, New York.

[39] Aashish Sheshadri and Matthew Lease. 2013. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*. 156–164. http://ir.ischool.utexas.edu/square/documents/sheshadri.pdf

[40] Pao Siangliulue, Joel Chan, Bernd Huber, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Self-sustainable Idea Generation in Creative Online Communities. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion (CSCW '16 Companion)*. ACM, New York, NY, USA, 98–101. https://doi.org/10.1145/2818052.2874335

[41] David W Stephens and John R Krebs. 1986. *Foraging theory*. Princeton University Press.

[42] Trevor Strohman, W. Bruce Croft, and David Jensen. 2007. Recommending Citations for Academic Papers. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*. ACM, New York, NY, USA, 705–706. https://doi.org/10.1145/1277741.1277868

[43] Yalin Sun, Pengxiang Cheng, Shengwei Wang, Hao Lyu, Matthew Lease, Iain Marshall, and Byron C. Wallace. 2016. Crowdsourcing Information Extraction for Biomedical Systematic Reviews. In *4th AAAI Conference on Human Computation and Crowdsourcing (HCOMP): Works-in-Progress Track*. http://arxiv.org/abs/1609.01017 3 pages. arXiv:1609.01017.

[44] Swaroop Vattam, Bryan Wiltgen, Michael Helms, Ashok K. Goel, and Jeannette Yen. 2011. DANE: Fostering Creativity in and through Biologically Inspired Design. In *Design Creativity 2010*. http://link.springer.com/chapter/10.1007/978-0-85729-224-7_16

[45] S. Wuchty, B. F. Jones, and B. Uzzi. 2007. The increasing dominance of teams in production of knowledge. *Science* 316, 5827 (2007), 1036–1039.

[46] James Zou, Kamalika Chaudhuri, and Adam Kalai. 2015. Crowdsourcing Feature Discovery via Adaptively Chosen Comparisons. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Received April 2018; revised July 2018; accepted August 2018