Towards a Virtual Librarian for Biologically Inspired Design

Ashok Goel, Kaylin Hagopian, Shimin Zhang, and Spencer Rugaber

Abstract In biologically inspired design, designers typically search for natural language documents describing biological systems relevant to their problems. Then they construct an understanding of the biological systems described in the documents for transfer to a given problem. These are difficult, labor intensive and time consuming processes. Thus, we are constructing a virtual librarian called IBID for supporting designers in locating and understanding biology articles relevant to their design problems. IBID first extracts knowledge of the function, the structure, and portions of the causal mechanisms of biological systems from their natural language descriptions. Then, it organizes this knowledge as a Structure-Behavior-Function (SBF) model. Finally, it uses the SBF annotations to retrieve biology articles relevant to design queries. To extract causal mechanisms, IBID uses machine learning techniques to identify portions of a document that describe causal processes.

1 Introduction

Biologically inspired design is a well-known paradigm that uses nature as a source of practical, efficient and sustainable designs to stimulate design of technological systems [1, 2]. However, not all architects, engineers, and designers are experts at biology [3]. Thus, not all designers have knowledge of a large number of biological systems stored in their internal memories, or a deep understanding of the biological systems available in their memories. Instead, in practice, given a design problem, many designers search for natural language documents describing biological systems and then construct an understanding of the retrieved systems for potential transfer to the design problem. In [4], we found that often this process is a labor intensive and time consuming because of the difficulty of finding articles relevant to a design problem, recognizing the relevance of a biological article to a design

A. Goel (✉) · K. Hagopian · S. Zhang · S. Rugaber
Design and Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA 30308, USA
e-mail: goel@cc.gatech.edu

J. S. Gero (ed.), *Design Computing and Cognition'20*,
https://doi.org/10.1007/978-3-030-90625-2_21

problem, and understanding the biological article deeply enough to enable analogical transfer of the biological knowledge to the design problem.

In earlier times, a human librarian might help designers in navigating a physical library and locating biology articles relevant to their design problems. In the current era, when increasingly large amount of biology literature is available online, there is a similar need for *virtual librarians* for finding relevant biology articles. In addition to locating relevant articles, we also want the virtual librarian to help designers in constructing a deep enough understanding of the biological systems described in an article to support analogical transfer.

In this paper, we briefly outline the IBID interactive system for supporting designers in locating and understanding biology articles relevant to their design problems. IBID (for Interactive Biologically Inspired Design) first extracts knowledge of the function, structure, and causal mechanisms of biological systems from their natural language descriptions. Then, it organizes this knowledge as a Structure-Behavior-Function (SBF) model [5]. Finally, it uses the SBF annotations to retrieve biology articles relevant to new design queries. Below we first present the conceptual design of IBID. Then, we briefly describe IBID's knowledge-based method for extracting portions of a causal process from text. A fully automated solution to the general problem of extracting causal processes from text is not yet available. Thus, next we describe IBID's use of statistical machine learning techniques to identify portions of a document that describe causal processes. The identified portions can potentially be examined by the human designer for deeper analysis.

2 Conceptual Architecture of IBID

Let us consider an expert designer who, like many designers, is a novice in biology. Let us suppose that the designer is interested in designing a system for transporting water to remote regions in her country. Given a large corpus of biology articles, how may we design a virtual librarian to help the designer locate and understand biology articles relevant to the design problem?

There are three stages in building a virtual librarian, starting with knowledge representation. AI research has developed precise languages for *representing* many kinds of knowledge. In the context of the current work, Julian Vincent in the United Kingdom has developed a detailed language for capturing knowledge of biological systems [6]. Our research laboratory has developed a more abstract language called Structure-Behavior-Function (SBF; [5, 7]) for expressing design problems as well as design patterns and principles. Briefly, an SBF model of a biological or a technological system explicitly specifies the structure of the system (the components and the connections among them), the functions of the system (the outcomes of the system), and the causal mechanisms (the system's behaviors) that explain how the structures of the system achieve its functions. The SBF model derives from Chandrasekaran's Functional Representation scheme [8, 9] and are similar to but distinct from the Function-Behavior-Structure model [10] and Function-Behavior-State model [11].

In previous work [12], we found that SBF models help a designer understand a complex biological system more deeply so as to better answer questions about its functioning. This led us to develop the Biologue system [13] for manually annotating biology articles based on SBF models, accessing biology articles relevant to a design problem based on the annotations, and using the annotations to help the designer understand the article in terms of SBF models. Experiments with Biologue showed that if biology articles are annotated with SBF models of the biological systems described in the articles, then many designers are better able to both locate biology articles relevant to their design problem and understand how the biological systems work. Thus, we posit that it may be productive if the virtual librarian for biologically inspired design too uses the SBF model as the knowledge representation scheme for capturing knowledge of biological systems.

Second, the design of a virtual librarian requires a scheme for *organizing* knowledge of biological systems in a digital library and methods for *accessing* the knowledge as and when needed. AI research has developed detailed schemes for organizing knowledge into digital libraries and accessing it from the libraries. In the context of the present work, The Biomimicry Institute has developed a library called AskNature ([14]; https://asknature.org/) containing hundreds of biological strategies and their application to biomimetic design. We have developed a digital library called DANE ([15]; http://dilab.cc.gatech.edu/dane/) that captures an understanding of a biological system in the SBF language, as well as a digital library named DSL of searchable natural language documents describing case studies of biological inspired design [16]. These libraries enable a designer to locate a biological system relevant to a design problem; they also scaffold the designer's comprehension of the biological system. Our design of IBID is specifically targeted towards acquiring knowledge of DANE's SBF models.

Third, the design of a virtual librarian requires methods for automatically *acquiring* knowledge of biological systems or automating the process as much as possible. AI research has developed several computational methods for automatically or semi-automatically acquiring knowledge for populating digital libraries. For example, in the context of this work, AI researchers have developed methods for acquiring biological knowledge from natural language documents on the Web [17]); [18] summarizes early attempts to develop computational techniques for literature-based discovery in biologically inspired design. More recent efforts include use of AI techniques for accessing and classifying natural language documents describing biological systems [19, 20], and discovering structure in patent databases [21]. Mueller et al. [22] have proposed acquiring biological design knowledge directly from animal fossils through machine vision.

In past work, our laboratory developed a virtual librarian for biologically inspired design based on IBM's Watson tool: it accessed biology articles relevant to a design query and answered questions based on the retrieved articles [23].

The IBID system in the current discussion focuses on extracting SBF models of biological systems from natural language documents, and using the SBF annotations to locate biology relevant to a function specified in a controlled vocabulary. IBID (http://dilab.gatech.edu/ibid/) operates in two modes. First, it extracts SBF models of biology articles in a given corpus and annotates the articles with structural, behavioral and functional terms. Given a research article describing a biological system from a journal such as Chrispeels & Maurel [24] article in *Plant Physiology*, IBID extracts the function, the structure, and parts of the causal behaviors of the system. Second, given a design query, IBID locates biology articles relevant to the query based on the structural, behavioral and functional annotations.

Figure 1 shows the full functionality of IBID for its three use cases: (1) End users such as engineers and designers looking for biology articles relevant to their design problems, (2) Knowledge engineers extending IBID's knowledge representation vocabulary, and (3) System administrators adding to its repository of analyzed papers. Figure 1 also specifies the actions available to each user type; the arrows in the figure indicate progression of steps and/or access to/from the database.

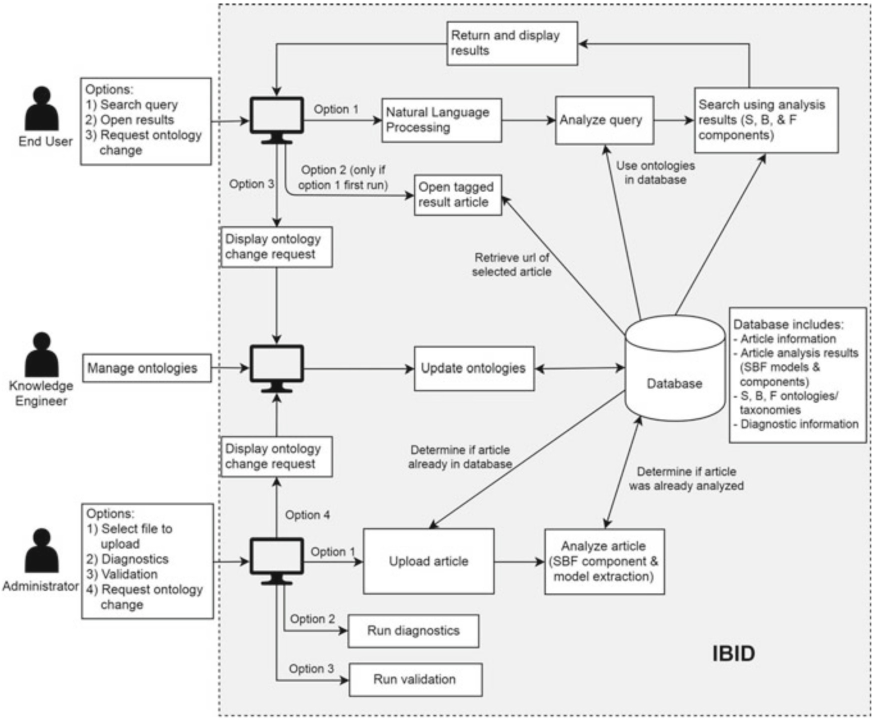
Return and display results
End User
Options:
1) Search query
2) Open results
3) Request ontology change
Option 1
Natural Language Processing
Analyze query
Search using analysis results (S, B, & F components)
Option 3
Option 2 (only if option 1 first run)
Open tagged result article
Use ontologies in database
Display ontology change request
Retrieve url of selected article
Knowledge Engineer
Manage ontologies
Update ontologies
Database
Database includes:
- Article information
- Article analysis results (SBF models & components)
- S, B, F ontologies/ taxonomies
- Diagnostic information
Display ontology change request
Determine if article already in database
Determine if article was already analyzed
Administrator
Options:
1) Select file to upload
2) Diagnostics
3) Validation
4) Request ontology change
Option 4
Option 1
Upload article
Analyze article (SBF component & model extraction)
Option 2
Run diagnostics
Option 3
Run validation
IBID

Fig. 1 The conceptual architecture of IBID

3 Extant Tools

Before we describe the conceptual architecture of IBID in more detail, we note that it uses several extant tools including the following:

3.1 Stanford Natural Language Parser

The Stanford Natural Language Parser (http://nlp.stanford.edu/software/lex-parser.shtml) generates parse trees of input sentences. IBID uses the core Stanford Natural Language Parser tool to construct the parse trees for sentences in biology articles (and natural language design queries). IBID uses the parse tree of a sentence to help identify if a part of the sentences refers to the structure, behavior or function of the biological system described in the article.

3.2 WordNet

WordNet (https://wordnet.princeton.edu) is a large lexical database of the English language in which different parts of speech are grouped into sets of synonyms (*synsets*), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. Given a design query expressed as an English language sentence, IBID uses WordNet to widen the set of search terms.

3.3 VerbNet

VerbNet (https://verbs.colorado.edu/~mpalmer/projects/verbnet.html) is an on-line verb lexicon that includes specific syntactic information and indications of verb class membership. Each verb class in VerbNet is described by its frames, thematic roles, and arguments. IBID uses VerbNet to identify and extract function terms from articles and expand its vocabulary of functions.

3.4 Vincent's Vocabulary for Structure of Biological Structures

Julian Vincent has developed a detailed vocabulary for describing the structure of biological systems (Vincent [6]). IBID uses a small part of his vocabulary as a domain-specific controlled vocabulary of biological structural components.

3.5 Domain-Independent Vocabularies for Structure, Behavior, and Function

We have developed domain-independent vocabularies of the structure, the behaviors, and the function of complex systems that build in part on the extant SBF vocabulary (Goel et al. [5]). Rugaber et al. [25] describes IBID's functional vocabulary. IBID uses these controlled vocabularies to capture the structural, behavioral, and functional concepts and relationships in the description of a biological system and in design queries.

4 Extraction of Structure, Behavior and Function from Text

The current version of IBID extracts functions, structure, and parts of the causal behaviors of a system from its natural language description. For each sentence in a biology article, IBID uses the Stanford NLP parser to obtain its phrase structure grammar representation in the form of a tree. Each valid phrase's start token in the tree represents the root node of a subtree whose leaf words are combined to create a logical sentence component. For example, one component of "Minute water droplets from fog gather on its wings; there the droplets stick to…" is "Minute water droplets from the fog gather on its wings".

4.1 Function Extraction

As indicated above, IBID uses a domain-independent controlled vocabulary of function terms [25]. Each term in this controlled vocabulary is expressed as a frame in VerbNet. The first step of function analysis is to generate a Stanford Dependency (SD) object for a given sentence component, the root of the SD tree is the predicate of the sentence and is then stemmed to produce the root verb. For example, "gather" is the root verb for the component "Minute water droplets from the fog gather on its wings". SD also provides information on whether the root verb is passive by listing any passive nominal subjects. Root verbs for which there are VerbNet records will have their VerbNet syntactic frames matched against the parser's Part-Of-Speech (POS) tags. For others, a custom algorithm uses WordNet to find the closest matching VerbNet word. The best matched predicate, its VerbNet syntactic frame, thematic relations mapping from our sentence component to the frame, and the sentence itself are saved in the database. IBID annotates the article with all this functional information.

4.2 Behavior Extraction

We have developed a domain-independent vocabulary for behavioral concepts and relations. Each behavioral term in this controlled vocabulary is expressed as a frame in VerbNet. As in function extraction, each sentence component is parsed for its predicate and then stemmed. Next, the root verbs are matched against causal verbs in our vocabulary of behavioral terms. If the system finds a causal verb, then IBID replaces it with a verb token and matches it against a list of predefined regular expressions capturing various forms of causal patterns. These causal regular expression patterns also delineate the sentence component's cause and effect clauses. The causality record, which includes a stemmed predicate, its cause/effect clauses, and the original sentence component are then saved in the database. Finally, IBID annotates the article with this behavioral information.

4.3 Structure Extraction

In [26], we describe IBID's extraction of structure of a biological system from its textual system. Briefly, IBID searches each sentence in the biology article for terms in Vincent's domain-dependent structural vocabulary. If it identifies a structural term, it then searches for adjectives that describe the structure/nouns. In addition, IBID uses WordNet to find synonyms, hyponyms, and meronyms for each structural term identified. IBID performs this additional search to map the structural terms from the domain-specific vocabulary into our domain-independent structure vocabulary. IBID then annotates the article with all this structural information.

5 Search

IBID is an interactive system intended to support complex human-AI interaction. Thus, it uses two kinds of search to locate biology articles in a corpus relevant to a design query: faceted search in which it uses domain-independent controlled vocabularies of structure, behavior and function terms; and search based on design queries stated as English language sentences. In the latter case, the current version of IBID does not yet use behavioral knowledge for locating biology articles.

Faceted search is based on a controlled vocabulary for function, behavior and structure. The function facet's controlled vocabulary has eight high-level function terms and multiple sub-level terms described in [25]. Given a designer's selection of functional terms, IBID searches the functional annotations on articles for the selected verbs. The behavior facet's controlled vocabulary is made up of a list of causal verbs. When a designer selects causal verbs of interest, IBID searches for articles possessing one or more behavior annotations with the selected verbs (or synonyms) as their foci. Cheong & Shu [27] describe a similar method for extracting causally-related functions. In IBID, the functional and behavioral terms derive from the SBF model and are also related to structural terms. Perhaps more importantly, these knowledge-based methods are able to extract only parts of the causal process. Thus, IBID seeks to combine them with machine learning techniques as described below. The structural facet's controlled vocabulary consists of the domain-independent structural vocabulary we have developed. Recall that when IBID extracts structural terms from biology articles, the terms are domain-dependent. In the current version of IBID, we manually map the domain-dependent structural annotations on the biology articles and the domain-independent terms in the faceted search. We intend to automate this process (and IBID already performs automated ontology alignment for natural language search).

In addition to faceted search, IBID can search based on design queries expressed in English sentences. Consider the query:

```
"I want to create a system for transporting liquid."
```

IBID first identifies functional and structural terms in the design query (verbs and nouns, respectively) and lemmatizes/stems them. For the given input, IBID finds *Function: [want, create, transport]* and *Structure: [system, liquid]*. Next, IBID enlarges this query by adding domain-independent structure terms and high-level function terms. For the given input, this results in *Function: [acquire, want, construct, create, move, transport]* and *Structure: [system, portion, liquid]*. Finally, IBID uses the same mechanism as in its faceted search based on the structural and functional annotations on the biology articles.

6 Preliminary Testing of Structural Queries

We have conducted preliminary evaluation of IBID's ability to retrieve biology articles based on structural design queries [26]. We begin with a piece of text taken from a biology research article [24] that we have used for evaluating parts of IBID:

> Bulk flow of water across a membrane occurs in response to an osmotic or hydrostatic gradient. Osmotic water permeability is readily measured in small vesicles or cells by the stopped-flow light-scattering technique, a method that relies on the dependence of light scattering on vesicle or cell volume, and is used to quantitate the time course of net water flow that occurs in response to transmembrane osmotic gradients. The osmotic gradients are established by adding an impermanent solute to the external solution. With the help of other chemical and physical methods to measure diffusional and osmotic water transport across biological membranes ...

We selected seven participants for our study, where the participants were not experts in biology (as with most biologically inspired designers). We asked the participants to list the structure terms in the above paragraph. Structure terms refer to the components, substances and connections of a system. For instance, in the following sentence: "Trees can transport water from the ground by their vascular system.", "water", and "vascular system" are the structure terms of the system "tree".

We gave exactly the same text and problem to IBID and compared the results with the human participants. We used the commonly used F_1 metric for the comparison, as it captures both the fraction of relevant terms that were retrieved as well as the fraction of the retrieved terms that were relevant. We found F_1 for identifying structural terms in the above experiment to be 79%. An interesting observation is the recall was higher than the precision, meaning that there were a larger number of false positives as compared to false negatives.

7 Machine Learning for Causal Relation Discrimination

Current natural language processing techniques for extracting causal processes work only in limited contexts (e.g., [28]); a general AI technique for automatically extracting causal process from a natural language document is not yet known. Thus, IBID's knowledge-based technique is able to extract only portions of the causal processes of a biological system from its natural language description. We posit that it would be useful to complement this top-down approach with a machine learning technique that performs bottom-up pre-processing on natural language documents to focus the knowledge-based extraction of causal behaviors on specific portions of a natural language document.

To focus the text analysis process so that only potentially relevant parts of a biological document are analyzed deeply, we have prototyped a causal biological process discriminator. Since our definition of a biological process (BP) has causality as one of its key elements, this system aims to classify paragraphs in an article that describes a causal relation involving at least one biological entity. The working definition of a biological process for the algorithm has three main components: (i) a biological organism or an entity closely related to a biological entity, (ii) one or more causal relations relating to at least one entity from the previous list, and (iii) a function served by the previous causal relations.

7.1 Approach

Our approach for causal relations discrimination has four main components. First, we apply term frequency inverse document frequency (TF-IDF) on a biology article to pick up important entities within the document. The algorithm then passes the top candidates to a knowledge database to filter for the topic biological entity candidates. The entities are then combined with causal adverbs and causal patterns to form a causal chain within an article section, which is finally passed through to a classifier to produce a BP classification.

7.2 Biological Entity Detection

The first step of analyzing a biological document is picking out keywords that are important to an article. For example, an article that describes the properties of spider silk, some of these keywords may include *spider*, *silk*, *fiber*, *spinneret*, and *strength*. In order to accomplish this, the document is compared against a non-biological domain text corpus via TF-IDF, a popular weighting scheme for words used by search engines and document query tools. The term frequency portion of the metric measures how important a particular word is to the document, while the inverse document frequency portion is used to filter out pronouns and other common stop-words not of interest for the analysis. In our case, Reuters 21,578 text categorization text collection is used as a benchmark. The result of the TF-IDF process is a list of words ordered by their importance to the article.

To differentiate words that are not just important to the document but also are related to biological entities, an outside knowledge source is required. This is accomplished by querying each word through DBpedia (https://wiki.dbpedia.org/), a database of structured content extracted from Wikipedia [29]. We leverage DBpedia's keyword search functionality to specifically search for entities that match both the keyword found and belong to the query class *species*. This ensures that only organism-related entities are saved, even if the keyword in question is not an entity. For example, in the previous example *spider* would return results on species of spiders, *silk* would return both spiders and silkworms and is retrained as a biology related entity while *strength* would not return any result and is discarded.

7.3 Causal Knowledge Graph Construction

After the key biology related entities are found, the individual sections of a document are analyzed and converted into a knowledge graph before serving the graph's key metrics as input to a classifier. To do this, each article paragraph is first tagged by their *part of speech* tags using the Python Natural Language Toolkit (NLTK; https://www.nltk.org/) library, then matched against a set of causal verbs and patterns as developed by Khoo et al. [30] for causality detection.

The causal graph creation is accomplished in a two-step manner. During the first pass, for each causality match found in a sentence, the cause and effect noun phrases are extracted if they are also one of the biological entities found in the previous section. The entities are then extracted and stored in a knowledge graph, with an edge connecting the two. In the second pass, each additional causal pattern matched is added to the knowledge graph if at least one of the two entities is already in the graph. This is repeated until no more nodes are added to the graph after an entire pass through the paragraph. The rationale for the two-step approach is to ensure all entities in the graph have a path to a biological entity, reducing the chance of experiment-related and otherwise non biology related entities being included during graph construction. This can result in the creation of one or more causality graphs. Finally, the total number of nodes in the knowledge graphs, the number of biological and non-biological entities, the total number of edges, the total number of edges linking biological entities and the size of the largest graph are outputted. Because most classifiers cannot take a graph model as input, our feature set is engineered to contain the pertinent information about the causal relation in our knowledge graph. The features are used to both train a classifier with human annotated labels as well as to serve the role of input at prediction time.

7.4 Discriminator Testing

To test the performance of the discriminator algorithm, a collection of 25 articles was hand annotated (based on the definition of a biological process provided earlier as well as their perceived usefulness for a designer/engineer looking for inspiration from biology) on a paragraph level to be used as a testing/validation set. The articles were selected from the existing IBID article database, each describing either a study conducted on some biological entity or is a general-purpose scientific article describing a biological process or a species. The 421 total annotated paragraphs from them were used as both a training and validation set to test a wide range of classifiers using the Python Scikit-Learn library. Results using an 80/20 training/validation set split and averaging all classifier results over 10 trials can be seen in Table 1. From the initial validation results, Gradient (Boosting) Trees were selected for testing on the testing set.

Table 1 Classifier performance on validation data

Classifier	Weighted accuracy	Precision	Recall	F-beta score
SVM	72.8%	0.81	0.72	0.76
Decision tree	70.9%	0.79	0.71	0.74
Gradient tree	86.4%	0.83	0.86	0.84
Random forest	76.5%	0.83	0.76	0.79
Logistic regression	69.1%	0.82	0.69	0.74

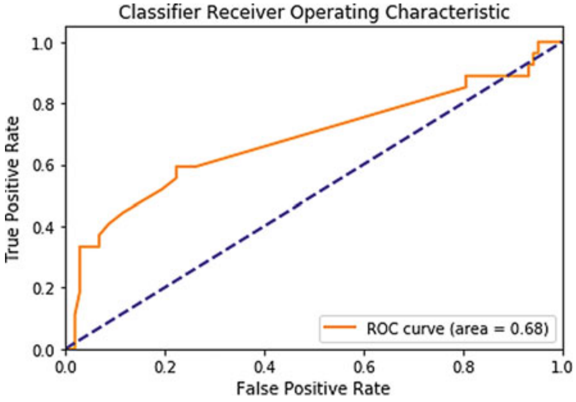

Fig. 2 ROC curve of algorithm performance on test data

As the paragraphs from both the training and validation data-sets came from the same set of articles, a new set of 5 articles was annotated in order to prevent the model from over-fitting. The new articles (130 paragraphs) were written by different authors and described different biological processes, which prevents the classifier from capturing the writing style of the training set authors and mimics the distribution of articles likely seen by an IBID-like tool in practice. The articles cover a range of topics, ranging from female frog calls to the vision system of cormorants. Over the new set of paragraphs, the classifier performed a little worse than it did on the previous test, achieving an accuracy of 81.5% with precision of 0.793, recall of 0.812 and F-beta Score of 0.777.

To get a better understanding of classifier performance, the Receiver Operating Characteristic plot was generated to see the effect of varying the discrimination threshold, as we are more interested in lowering our rate of false negatives than false positives. From Fig. 2, we see that to increase our true positive rate from 0.6 to 0.8 would require a corresponding increase of false positives from 0.25 to 0.65, indicating the lowering of our classifier's threshold may not be an effective method for reducing false negatives in our predictions.

In order to further study the characteristics of the algorithm output, two additional papers were annotated, and classifier outputs studied in-depth. One of the most confident false positives was produced by the classifier on the following paragraph from [31]:

> Despite the high abundance of mucus and mucus aggregates in coral reefs, their development over time in the reef water has not been investigated. An understanding of temporal changes in mucus aggregate composition is critical, because these aggregates potentially have an important function in the transfer of energy from the corals to other reef organisms, and in the trapping of organic matter and nutrients from water over passing the reef. The objectives of this study were (1) to assess whether dissolved mucus can cause mucus-particle aggregates, and (2) ...

Here the paragraph describes the central biological process investigated by the paper, that is, whether coral mucus can attract organic matters in the shallow coral reefs and form a key component of the nutrient cycle. The paragraph also contained key causal phrases such as *cause*, *release*, *assess*, and *trapping* as well as the key biological entities studied (*coral, mucus, reef*). However, the paragraph was not initially labelled as the description of a biological process because it does not talk about the production process of coral mucus and appears to be speculative in nature.

On the other hand, the following is an excerpt from [32] of the most confident false negative paragraph predicted:

> The improvements in local buckling resistance under axial load or bending moment are mixed, with most species achieving some improvement in local buckling moment resistance (Fig. 8(b)). ... The short spines of the hedgehog (Erinaceus) and the spiny rat (Hemieehlnus) are required to act as shock absorbers as much as armour and protection to discourage predators, hence the high structural efficiency requirement and the need to delay local buckling until the internal stresses have almost reached ...

In this example, the biological process is described in a fleeting manner (spines acting as shock absorbers and armour). The causal relation is described in an indirect manner, (act as shock absorbers) instead of ‘spines absorb shock’ and ‘act … as much as amour and protection’ instead of ‘protect against predators’. Because the biological process is mentioned both briefly and in an indirect manner, it was not successfully picked up as a part of the knowledge graph and ultimately classified incorrectly with a very high confidence level.

8 Discussion

AI scientists have long dreamt of supporting human creativity, for example, creativity in scientific discovery and technical design. As early as the mid 1960s, just a decade after McCarthy and colleagues [33] first coined the term "artificial intelligence", Feigenbaum and colleagues developed the DENDRAL knowledge system for abducing the structures of chemical molecules from mass spectroscopy data [34]. In the 1970s and the 1980s, knowledge systems such as R1 [35], AIR-CYL [36], PRIDE [37], VEXED [38] and VT [39] sought to capture design expertise in the form of design concepts, rules, constraints and plans. The hope at the time was that if AI systems could help capture large-scale expert knowledge, then the systems could support problem solving, design, discovery, and creativity at scale. However, as is well known now, experts at a given task in a given domain do not always agree, much of expert knowledge is tacit, it is difficult to elicit knowledge from experts, and it is also difficult to maintain it over time.

In the second wave of AI research on supporting creativity in the 1980s and1990s, the focus expanded from knowledge to include experience. While expert knowledge may be tacit and difficult to elicit, the argument went, experts have external representations of their experiences: For example, most designers develop design briefs and many scientists keep research journals. If AI systems could capture these experiences in the form of case libraries, the systems could support problem solving, design, discovery, and creativity at scale. CYCLOPS [40], STRUPLES [41], ARGO [42], ARCHIE [43] and AskJef [44] were among the first case-based systems for supporting design creativity. While these interactive systems provided annotated digital libraries of design cases, they left the task of case adaptation to the designer. Although research in this paradigm continues, once again it has been difficult to acquire large libraries of well-documented cases, as well as difficult to maintain them over time.

In this century, the notion of literature-based discovery has given rise to a third wave of AI research on supporting creativity. Literature-based discovery analyzes publicly available scientific literature to find connections among seemingly distant entities and analogies between seemingly different relationships and processes [45, 46]. Bruza & Weeber [47] compile an anthology of work on literature-based discovery; Henry & Mcinnes [48] provide a recent survey. The hope is that publicly available literature can ameliorate some of the difficulties of earlier AI attempts at creativity. In the context of AI research on creativity, Abgaz et al. [49] use natural language processing to find analogies between constructs in research papers on computer graphics, and Lavrac et al. [50] describe text mining techniques for detecting bridging concepts between seemingly unrelated terms such as *migraine* and *magnesium*.

One important issue in AI research on literature-based discovery and design is the balance between the techniques of knowledge-based reasoning and statistical machine learning. Previous research has ranged from using mostly machine learning techniques [20, 50] for classification documents, to using mostly knowledge-based techniques [17, 49] for analogical retrieval and mapping. As we described in the previous section, IBID combines knowledge-based reasoning and statistical machine learning: While IBID's conceptual architecture is knowledge-based, it uses machine learning to accomplish specific tasks defined by its architecture. The results of the machine learning technique can then be processed by the knowledge-based method or by the human designer.

9 Conclusions

IBID is an interactive AI system for helping biologically inspired designers locate and understand biology articles that describe biological systems relevant to a design query. Given a design problem, many designers typically search online for biology articles for inspiration. The analogical retrieval, mapping and transfer from biology to design often is mediated by functional models of the biological systems described in the articles. Thus, IBID first extracts structural, behavioral and functional terms in biology articles and annotates the articles with the terms. Then, given a design query, IBID locates the biology articles relevant to the query based on the articles' annotations. IBID uses two kinds of search to locate biology articles: faceted search based on domain-independent controlled vocabularies of structures, behaviors and functions; and natural language query search for function and structure.

The problem of extracting complex elements, such as the behaviors of a biological system in the form of casual processes, from natural language documents remains unsolved in AI. This problem had confounded our preliminary work on the IBID project. This work posits that addressing this problem requires a combination of knowledge-based and machine learning techniques. In general, techniques of statistical machine learning are computationally expensive, and require large amounts of labeled data but do not perform deep semantic analysis. A knowledge-based architecture for the overall task can help focus machine learning on specific subtasks thereby controlling the computational cost. The machine learning techniques can, then, make use of standardized datasets to produce preliminary results in the form of portions of natural language documents for further semantic analysis. Our experiments with the IBID system explore this approach. Thus, as described in this paper, IBID's conceptual architecture spawns the subtask of behavior extraction and focuses the machine learning techniques to the behavioral processes. In addition, its SBF models provides the behavioral terms for searching the biology articles. It then uses machine learning techniques to classify specific portions of natural language documents that specify biological processes. While the machine learning techniques do not provide any guarantee of correctness, their output can be further analyzed for behavior extraction either by the knowledge-based methods or the designer.

Acknowledgements We are grateful to Julian Vincent for sharing his domain-specific ontology for describing the structure of biological systems (Vincent [6]). We thank Swapnal Acharya, Kimisha Mody, Pablo Boserman, Daniel Diaz, and Ruth Petit-Bois for their contributions to this research.

References

1. Benyus J (1997) Biomimicry: innovation inspired by nature. William Morrow
2. Vincent J, Mann D (2002) Systematic technology transfer from biology to engineering. Phil Trans Roy Soc London A Math Phys Eng Sci 360(1791):159–173
3. Yen J, Helms M, Goel A, Tovey C, Weissburg M (2014) Adaptive evolution of teaching practices in biologically inspired design. In: Goel A, McAdams D, Stone R (eds) Biologically inspired design: computational methods and tools. Springer, London: pp 153–200
4. Vattam S, Goel A (2013) Seeking bioinspiration outline: a descriptive account. In: Proceedings 19th international conference on engineering design (ICED13), Seoul, Korea, August 2013, pp 517–526
5. Goel A, Rugaber S, Vattam S (2009) Structure, behavior and function models of complex systems: the structure-behavior-function modeling language. AIEDAM 23:23–35
6. Vincent J (2014) An ontology of biomimetics. In: Goel A, McAdams D, Stone R (eds) Biologically inspired design: computational methods and tools. Springer, London, pp 269–285
7. Goel A (2013) One thirty year long case study; fifteen principles: implications of an AI methodology for functional modeling. AIEDAM 27(3):203–215
8. Chandrasekaran B (1994) Functional representation: a brief historical perspective. Appl Artif Intell 8(2):173–197
9. Chandrasekaran B, Goel A, Iwasaki Y (1993) Functional representation as design rationale. IEEE Comput: 48–56
10. Gero J (1990) Design prototypes: a knowledge representation schema for design. AI Mag 11(4)
11. Umeda Y, Tomiyama T (1997) Functional reasoning in design. IEEE Expert 12(2):42–48
12. Helms M, Vattam S, Goel A (2010) The effects of functional modeling on understanding complex biological systems. In: Proceedings 2010 ASME conference on design theory and methods, Montreal, Canada, August 2010
13. Vattam S, Goel A (2011) Foraging for inspiration: understanding and supporting the information seeking practices of biologically inspired designers. In: Proceedings 2011 ASME DETC conference on design theory and methods, Washington DC, August 2011
14. Deldin J, Schuknecht M (2014) The AskNature database: enabling solutions in biomimetic design. In: Goel A, McAdams D, Stone R (eds) Biologically inspired design: computational methods and tools. Springer, London: pp 17–27
15. Goel A, Vattam S, Wiltgen B, Helms M (2012) Cognitive, collaborative, conceptual and creative - four characteristics of the next generation of knowledge-based CAD systems: a study in biologically inspired design. Comput Aided Des 44(10):879–900
16. Goel A, Zhang G, Wiltgen B, Zhang Y, Vattam S, Yen J (2015) On the benefits of digital libraries of analogical design: documentation, access, analysis and learning. AIEDAM 29(2)
17. Chiu I, Shu L (2007) Biomimetic design through natural language analysis to facilitate cross-domain information retrieval. AIEDAM 21(1):45–59
18. Shu L (2010) A Natural-language approach to biomimetic design. AIEDAM 24:507–519
19. Kruiper R, Vincent J, Chen-Burger J, Desmulliez M (2017) Towards identifying biological research articles in computer-aided biomimetics. In: Proceedings conference on biomimetic and biohybrid systems. Springer: 242–254
20. Vandevenne D, Verhaegen P, Dewulf S, Duflou J (2016) SEABIRD: scalable search for systematic biologically inspired design. AIEDAM 30(01):78–95

21. Fu K, Cagan J, Kotovesky K, Wood K (2013) Discovering structure in design databases through functional and surface based mapping. ASME J Mech Des 135
22. Mueller R et al (2018) Biodiversifying bioinspiration. Bioinspiration & Biomimetics 13(5)
23. Goel A et al (2016) Using watson for supporting design creativity. In: Proceedings fourth international conference on design creativity, Atlanta, Georgia, October 2016
24. Chrispeels M, Maurel C (1994) Aquaporins: the molecular basis of facilitated water movement through living plan cells? Plant Physiol 105(1):9–13
25. Rugaber S et al (2016) Knowledge extraction and annotation for cross-domain textual case-based reasoning in biologically inspired design. In: Proceedings 24th international conference in case-based reasoning (ICCBR 2016), Atlanta, USA, October 2016: pp 342–355
26. Goel A, Acharya S, Mody K, Hagopian K, Zhang S, Rugaber S (2018) Extracting structural knowledge from natural language documents to support biologically inspired design. In: Proceedings AAAI fall symposium on AI and natural systems. https://www.grc.nasa.gov/vine/presentations-and-papers/
27. Cheong H, Shu L (2014) Retrieving causally related functions from natural-language text for biomimetic design. ASME J Mech Des 136(8)
28. Berant J, Srikumar V, Chen P-C, Huang B, Manning C, Linden A, Harding B (2014) Modeling biological processes for reading comprehension. In: Proceedings 2014 conference on empirical methods in natural language processing (EMNLP 2014), pp 1499–1510
29. Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) DBpedia-a crystallization point for the web of data. Web Semantics Sci Serv Agents World Wide Web 7(3):154–165
30. Khoo C, Kornfilt J, Oddy R, Myaeng S-H (1998) Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. Literary Linguist. Comput. 13(4):177–186
31. Huettel M, Wild C, Gonelli S (2006) Mucus trap in coral reefs: formation and temporal evolution of particle aggregates caused by coral mucus. Mar Ecol Prog Ser 307:69–84
32. Karam G, Gibson L (1994) Biomimicking of animal quills and plant stems: natural cylindrical shells with foam cores. Mater Sci Eng C 2(1–2):113–132
33. McCarthy J, Minsky M, Rochester N, Shannon C (1955) A proposal for a dartmouth summer project on artificial intelligence. www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html
34. Lindsay R, Buchanan B, Feigenbaum E, Lederberg J (1980) Applications of artificial intelligence for organic chemistry: the dendral project. McGraw-Hill Book Company
35. McDermott J (1982) R1: a rule-based configurer of computer systems. Artif Intell 19(1):39–88
36. Brown, D., & Chandrasekaran, B. (1989) *Design problem solving: Knowledge structures and control strategies*. San Mateo (Calif.): Morgan Kaufmann.
37. Mittal S, Dym C, Morjaria M (1986) PRIDE: an expert system for the design of paper handling systems. IEEE Comput 19(7):102–114
38. Steinberg L (1987) Design as refinement plus constraint propagation: the VEXED experience. In: Proceedings national conference on AI (AAAI-87), July 1987, Seattle
39. Marcus S, Stout J, McDermott J (1988) VT: an expert elevator designer that uses knowledge-based backtracking. AI Mag 9(1):95–111
40. Navinchandra D (1991) Exploration and Innovation in Design. Springer, New York
41. Zhao F, Maher M (1988) Using analogical reasoning to design buildings. Eng Comput 4:107–122
42. Huhns M, Acosta E (1988) Argo: a system for design by analogy. IEEE Expert 3(3):53–68
43. Barber J et al (1992) AskJef: integration of case-based and multimedia technologies for interface design support. In: Proceedings second international conference on artificial intelligence in design. Dordrecht: Kluwer: p 457–74
44. Pearce M, Goel A, Kolodner J, Zimring C, Sentosa L, Billington R (1992) Case-based decision support: a case study in architectural design. IEEE Expert 7(5):14–20

45. Swanson D (1986) (1986) Undiscovered public knowledge. Libr Q 56(2):103–118
46. Swanson D (2008) Literature-based discovery? The very idea. In: Bruza P, Weeber M (eds) Literature-Based Discovery. Springer
47. Bruza P, Weeber M (eds) (2008) Literature-Based Discovery, Springer
48. Henry S, Mcinnes B (2017) Literature-based discovery: models, methods, and trends. J Biomed Inf 74
49. Abgaz Y, Chaudhry E, O'Donoghue D, Hurley D, Zhang J (2017) Characteristics of pro-c analogies and blends between research publications. In: Proceedings 8th international conference on computational creativity, pp 1–8
50. Lavrac N (2019) Bisociative knowledge discovery for cross-domain literature mining. In: Veale T, Cardoso A (eds) Computational Creativity. Springer