Data Driven Techniques for Organizing Scientific Articles Relevant to Biomimicry

Yuanshuo Zhao,* **Ioana Baldini,†** **Prasanna Sattigeri,†**

Inkit Padhi,† **Yoong Keok Lee†** and **Ethan Smith‡**

*Georgia Institute of Technology
†IBM Research AI
‡Biomimicry Institute
yzhao314@gatech.edu, {ioana,psattig}@us.ibm.com, inkpad@ibm.com, yklee@us.ibm.com, ethan.smith@biomimicry.org

Abstract

Life on earth presents elegant solutions to many of the challenges innovators and entrepreneurs across disciplines face every day. To facilitate innovations inspired by nature, there is an emerging need for systems that bring relevant biological information to this application-oriented market. In this paper, we discuss our approach to assembling a system that uses machine learning techniques to assess a scientific article's potential usefulness to innovators, and classifies these articles in a way that helps innovators find information relevant to the challenges they are attempting to solve.

Introduction

Nature presents solutions to many of the challenges innovators and entrepreneurs face every day, and a growing number of these people have begun studying living systems as part of their efforts to solve human problems via an array of interrelated disciplines–including but not limited to biomimicry, biomimetics, and bionics-collectively referred to as biologically inspired design (BID) (Benyus 1997). This approach has helped a range of teams develop novel products, services, and systems (Smith et al. 2015) that are often more elegant, efficient, or economic than the status quo. An increased level of interest in BID is evident via an elevated number of academic papers on a wide variety of topics, a growth in patented inventions, and the blossoming of dozens of related consultancies, consortia, accredited academic programs, and conferences across industries around the globe over the past two decades. The elevation in academic output is indicative of overall research activity and interest in both the theory and practice of BID, the increased number of patents including biologically inspired components suggests that this approach helps lead to commercially viable solutions, and the grassroots evolution of related networks and services demonstrates a general interest in and appetite for BID.

As general interest in BID has grown, so too has research and development of ways to systematize processes, methods, and tools that facilitate access to and utilization of biological information that is potentially relevant to any given problem solver. But the processes, methods, and tools that have been developed to date do not yet meet the needs of these innovators to systematically and effectively leverage the volume of biological knowledge that the scientific community has developed over time. Many of the biologically inspired designs that the general public is familiar with have either been the products of serendipitous observation and insight (the inventor of Velcro claimed to have been inspired after observing the ability of the cocklebur to attach to the fur of his dog) or side-effects of focused study on a particular living system (Gecko tape structural adhesives on the market are a side-product of intensive research on geckos ability to walk on a variety of surfaces and orientations). Both of these methods are difficult to replicate at scale, as they either require an unpredictable eureka moment, or a significant amount of devoted research that may not directly relate to an application-oriented goal.

The bulk of recent research efforts to systematize BID have thus focused on processes, methods, and tools that support a third type of idea generation: cross-domain search and knowledge transfer. Sometimes referred to as Design-by-Analogy (Moreno Grandas et al. 2015), this is a replicable problem solving process in which attributes, relations, and purposes from a source problem or situation are mapped to an existing target solution or situation, often from a different domain than the source. For example, an engineer tasked with developing a stronger and lighter structural material suited to cold environments might create a list of desired attributes and applicable situations and then use those parameters to search through a number of domains for potential analogues. In the case of BID, the engineer might become generally familiar with biological organisms that produce strong and light materials with similar characteristics to those desired, and particularly focus on those that have evolved to do so in cold regions of the planet.

To be valuable at scale, BID tools should allow innovators to quickly and accurately find potential analogues, that are described using domain-agnostic structures and language that can be effectively translated back to the source problem. Over the past decade a number of teams, primarily from academia, have developed and researched an array of proof-of-concept tools to help facilitate various aspects of this approach.Of these tools, the Biomimicry Institute's AskNature (asknature.org) database, originally launched in 2008, has persisted as the largest and most-utilized offering to date,

drawing visitors with over 1,600 biological strategy articles describing how living systems meet the functional challenges presented by their environments. The bulk of these articles on AskNature provide a plain-language summary of one or more biological phenomena, reference one or more peer-reviewed articles or books, and offer an array of supplemental information for users wishing to learn more. At its current scale, AskNature serves as an interesting proof of concept, but given the millions of biological phenomena that have been observed and documented in books, journals, and online resources, the scale of its catalog seems insignificant at best. To date, AskNatures articles have been researched and created by trained specialists via a time-consuming and completely manual process that cannot be effectively scaled to meet the needs of the growing base of BID students and practitioners.

In an effort to facilitate a higher rate of content generation and open the door to new BID use cases, the authors of this paper developed automated methods to detect articles that may be relevant to AskNature and to classify articles into the top level of the Biomimicry Taxonomy schema of functions.

Biomimicry Taxonomy

The Biomimicry Taxonomy is a three-tiered schema of functions developed by the Biomimicry Institute to act as an analogical bridge between biology and a variety of other domains (Hooker and Smith 2016). The primary level of this schema includes eight groups which ultimately break down into 168 individual functions. The primary taxa include ”move or stay put”, ”protect from physical harm”, ”maintain community”, ”modify”, ”make”, ”process information”, ”break down”, and ”get, store, or distribute resources”.

This schema can be used to navigate AskNature's content, which is organized according to this schema. In addition, given a design challenge, the schema can help innovators think about their challenge in terms of function in order to identify questions they might ask nature. For example, when an innovator's goal is to design a specific structure that can adapt to hot conditions, they can click on the function ”protect from physical harm”, and the website will direct the designer into a pool of content relevant to this function. The user can further narrow down through the function classes until they find what they want.

Related Work

Biomimicry Our work is related to three lines of biomimicry research. The first body of work draws inspiration from biological systems to devise solutions in other domains. BID has impacted numerous fields such as sustainable design (Benyus 1997), mechanical design (Raibert et al. 2008), optimization algorithms (Yang 2014), and, more notably, material engineering (Hawkes et al. 2015). Although our work is not a bio-inspired invention per se, the techniques we developed can be included in an end-to-end ideation toold that accelerates such cross-disciplinary innovation. The second line of biomimicry research examines the processes behind bio-inspired problem solving to propose methodologies for transferring knowledge in biological systems. (Gentner 1983) and (Goel 1997) helped pioneer structure-oriented design-by-analogy, laying the foundation of many frameworks today. Subsequent research efforts conducted in-depth studies on the role of analogy in specialized domains such as product ideation (Dahl and Moreau 2002), engineering design (Helms, Vattam, and Goel 2009), and mechanical systems (Chakrabarti et al. 2005). Our work falls into the third line of research which develops tools to assist BID. There are a number of manually curated databases of biological knowledge, such as AskNature (Deldin and Schuknecht 2014), BioTRIZ (Vincent and Mann 2002), and DANE (Vattam et al. 2011). Most closely related to our work are systems capable of extracting knowledge from unstructured text such as scientific literature. However, existing systems rely on manually crafted rules to identify biological concepts even if knowledge-based or machine learning-based linguistic preprocessing modules may be used (Shu and Cheong 2014; Vandevenne et al. 2016; Kruiper et al. 2017). A recent system (Cheong et al. 2017) uses neural network word embeddings to compute similarity measures but it does not employ machine learning techniques to perform the core task of mining biological knowledge. In contrast, we employ word embeddings in supervised machine learning models to filter and identify biological functions. Such learning approaches can better cope with distilling patterns and handling ambiguity as we scale up to larger data sets (Jurafsky and Martin 2000; Witten et al. 2016).

Machine Learning Methods for Text Mining Our work exploits machine learning techniques for text mining. We utilize supervised learning (Joachims 1998; Berger, Pietra, and Pietra 1996; McCallum, Nigam, and others 1998; Liaw, Wiener, and others 2002) which was proven to be effective for numerous natural language processing tasks, such as text classification (Yang and Pedersen 1997), sentiment analysis (Pang, Lee, and others 2008), and analogy mining between pairs of products (Hope et al. 2017). Although there is a long history of research employing machine learning for natural language understanding, the problem of mining biological knowledge for BID has not been explored. There is a body of research in natural language that specializes in extracting bio-molecular terminologies (Kim et al. 2009; Nédellec et al. 2013). These research would be useful for preprocessing a subset of scientific publications but it is inadequate for BID since the taxomonies are not aligned. In addition, we also experiment with neural network models (LeCun, Bengio, and others 1995; Johnson and Zhang 2014) which have recently advanced the state-of-the-art for various tasks in natural language processing. The wealth of untapped machine learning methods presents an immense resource to bring biomimicry research to greater heights.

Dataset Curation

The success of machine learning and artificial intelligence (AI) algorithms highly depends on the quality of the available data. In the last decades, harnessing the power of the

crowd in collecting and cleaning data became one sure way of increasing the quality of the data used in machine learning algorithms. In this section, we describe our methodology for dataset curation and a crowdsourcing architecture that is scalable and easy to deploy.

Crowdsourcing task design

Our intent is to automatically or semi-automatically create the AskNature content such that AskNature can scale to the large number of resources available in the biology domain. For now, we restrict our focus to scientific articles, but eventually, we would like to include other content media types, such as videos and images. Consulting biology experts and biomimicry enthusiasts, we learned that for a biology paper to be a potential source of inspiration for a biomimicry solution, it needs to talk about a living organism, and, more specifically, it needs to describe a function that the organism performs and the mechanism through which the function is realized. Our long-term plan is to automatically identify organisms in papers and extract the portions of the text that identify the function and the corresponding mechanism. To be able to perform such a task, we need a considerable amount of training data to train and test machine learning algorithms. For the work presented in this paper, we focus on building categorization services that decide whether a scientific abstract indicates a potentially relevant paper to biomimicry, and, if so, what type of function it belongs to, according to the primary level of the biomimicry taxonomy. While this task is simpler than our long-term goal, it is still challenging and requires significant training data. The remainder of this section describes the data collection effort.

The first task was to identify journals with content highly relevant to biomimicry. Journal of Experimental Biology is such a collection of papers. Once we collected the abstracts of the papers, we devised a questionnaire that we would like to fill in for each of these papers. This questionnaire includes questions about the organism, function and mechanism described in the abstract. In addition, we are also interested in classifying the function described in the abstract in one of the eight classes that are included at the primary level of the biomimicry taxonomy.

Data Collection Framework

The crowdsourcing application is implemented using serverless technology offered by IBM Cloud Functions (IBM 2017). The serverless computing promises to eliminate the burden of managing computational resources by allowing the developer to focus solely on the logic of the application. It achieves this goal by shrinking the unit of deployment to a single function. As such, a serverless application consists of a collection of functions. The sole responsibility of the developer is to create these functions, register them with the serverless platforms and compose them into applications. Whenever these functions are invoked, the serverless platform allocates resources to execute the function code. The serverless platform is responsible for fault-tolerance and autoscaling, while users are charged only for the time their functions are running. Usually, serverless functions are triggered by events, such as a user pressing a button, a user sending an email, a code repository commit or a twitter on a particular topic. As such, servereless platforms are a great fit for event-driven applications.

To create our content source, we crawled the papers from the Journal of Experimental Biology since 2000, we extracted the titles, authors and abstracts of all papers and stored all this information in a JSON data store. The collection includes about 3.2K papers. This step was also implemented as an action on the serverless platform. The questionnaire was translated into an input form using one of the drag-and-drop tools freely available on the internet and customized it to fit our needs. Each form refers to a different paper with a specific title, authors and abstract. Several of our input fields are conditional on the answers provided by the user. If the paper is not relevant to biomimicry, we are not interested in any of the potential answers to the rest of the questionnaire. As such, the fields in the form are displayed conditionally depending on the answer that the user provides. In addition, we anticipate that several papers may be talking about several functions related to the same organism. As such, we include a button in our form that, when pressed, displays a series of additional questions referring to the same paper/abstract.

In principle, we would like to collect several entries for each paper. We keep track of papers that have been serviced before through a simple mechanism we refer to as "crowd-cache". Before rendering each form, we consult our crowd-cache to determine which paper to include in the current form. The strategy that we implemented for servicing paper is the following. We are splitting our papers in chunks and for each chunk of papers we look to collect four different versions of answers before moving and servicing the next chunk of papers. This way we ensure that we have multiple answers for a set of papers before potentially collecting information on all the papers. The size of the chunk determines how many unique entries are collected before a paper is serviced again to collect the same information for the paper, albeit from a different crow-worker. While in theory this simple mechanism helps collecting several entries per paper, in practice, due to the relatively low number of crowd-workers, we collected multiple entries per paper for a low fraction of the total papers serviced.

An excerpt from our example of form rendition is shown in Figure 1. While we were designing the form, we conducted simple user studies with our colleagues and incorporated their feedback in the form design. For example, one user reported that some papers were extremely complex and she felt she could be more productive if she could skip some of the papers and continue with the ones that required less effort. While skipping some documents runs the risk of not having enough information on some papers, we decided to include a *Skip* button because, ultimately, we are interested in keeping our users as engaged as possible since we would like them to fill in the information on a large number of papers. One option to make sure that these papers are not left behind with no information being collected, we could include the identifiers for these papers in our crowd-cache and inter-spread these papers among the forms created for some of the users that do not seem to skip papers, assuming that

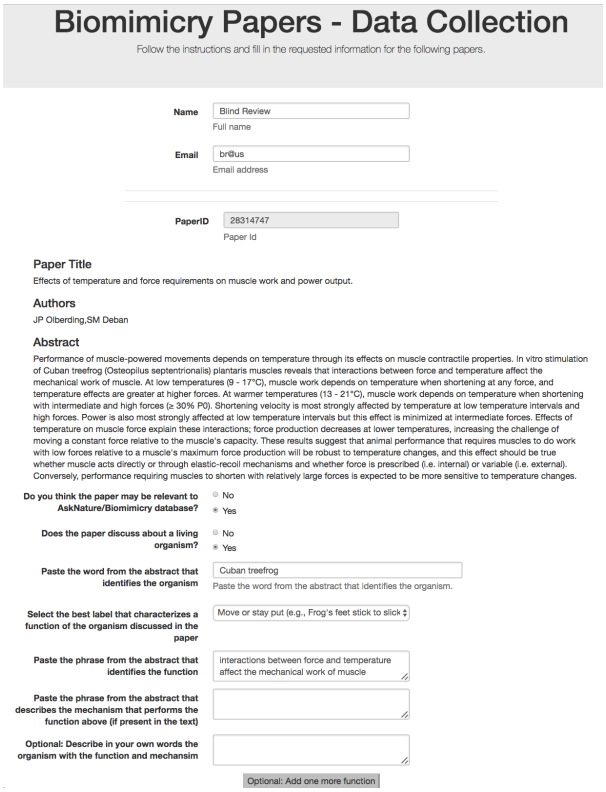

Figure 1: An excerpt of a sample form dynamically created and with conditionally displayed fields.

these users are the biology experts inclined to provide answers for any type of paper.

Once the form data is submitted, we perform some simple data cleaning. We first check whether the paper was deemed as relevant to biomimicry. If the paper was not identified as useful to biomimicry, we store only this information about the paper and ignore all other fields. If the paper was identified as relevant, we save all the answers that are non-empty and check whether optional answers are present, and if they are, we store them accordingly. All our data is stored in an Elasticsearch data store that is deployed in IBM Cloud.

Once the form is realized, the resulting HTML code is included in an action written in JavaScript. The title, abstract and authors are replaced with placeholder tokens. For each form rendition, these placeholders are respectively replaced with information retrieved from the database. For the two skip and submit buttons, we declare separate sequence actions which record the skipped article and clean the data, respectively, followed by regenerating the form for the next paper provided by the crowd-cache action. All actions are created and deployed in IBM Cloud Functions (IBM 2017).

Dataset Summary Statistics

Additional from the crowdsourcing collections, some of the data are scrapped from the web using Asknature's database, where the URL from Asknature's database is utilized to grab the articles if they are available on the internet. Around 130 articles have been scrapped. Together with the crowdsourcing data, the total dataset comprises of 844 unique abstracts. Each human annotator first determines whether an abstract is relevant. If an abstract is determined to be relevant, the annotator is further presented with an option to select one or more function classes. Each article is labeled by one person, except for a small portion (7%) of the articles that are labeled by more than one person. For the articles with multiple answers, majority vote was used to determine the article's category. The data collection resulted in 457 relevant articles and 387 irrelevant ones. There are 439 abstracts that have a function label. There is only one abstract annotated with the "Other" function class which we discard for our experiments. The frequency breakdown for each function class is shown in Table 1.

Categorization into Biomimicry Taxonomy

The goal of the proposed system, which we call Biomimicry Explorer hereafter, is to perform indexing of the documents conditioned on the relevance to biomimicry. It helps a user narrows down on relevant sections of the database faster. With inclusion of more supervised data, the final objective would be to create a summarized snippet of the relevant articles highlighting the key information that includes organism, function, and mechanism.

We first classify the articles into *relevant* vs *irrelevant*, then we classify the relevant articles into function classes that defined by Biomimicry Taxonomy. We then extract the features (in our case n-gram phrases) that have high weights during the classification and use that information to extract the sentences.

Experiment Setup

We evaluated a number of machine learning models for our relevance and function classification experiments. For each learning model, we randomly partition the whole data set into a training set (90%) and a held-out test set (10%). Our neural network model trains on 70% of the training set and uses the remaining 30% as the validation set to tune the number of training epochs with early stopping. Because there are rare function class, such as "Break down", we use stratified partition for function classification to maintain the same class distribution in the training and the test set. We report the mean predictive accuracy on the test set over five random restarts.

Machine Learning Models and Feature Representation

For both the tasks of relevance detection and function classification, we used machine learning classifiers which were trained using the dataset we collected. Several classification algorithms were studied and compared including Logistic Regression, Naive Bayes, Random Forest, Support Vector Machine. We tried both bag of words (BoW) and term frequency-inverse document frequency (tf-idf) separately as features to train models using the aforementioned algorithms. For Naive Bayes learning, we used Bernouli and

ID	Count	Function	Example
1	6	Break down	Fungi breakdown hydrocarbons in crude oil
2	88	Get, store, or distribute resources	Camel noses prevent water loss during exhalation
3	19	Maintain community	Sea Anemone and Clownfish benefitfrom mutualism
4	23	Make	Spider silk is assembled on demand
5	54	Modify	The scales of pine cones flex passively in response to changes in moisture levels
6	83	Move or stay put	Frog's feet stick to slick surfaces
7	49	Process information	Owl ears map sound in three dimensions
8	116	Protect from physical harm	Hairs reflect light and dissipate heat to keep ants cool
9	1	Others	

Table 1: Frequency distribution of function types in our dataset.

Multinomial generative models and for random forest classifier, we applied gini impurity as criteria and trained them using both BoW and tf-idf features. We also employed Convolutional Neural Network (CNN) for the relevance classification problem. The architecture is based on the approach in (Kim 2014) and uses the pre-trained GloVe word embeddings (Pennington, Socher, and Manning 2014). Deep learning based models have great potential when we have substantial amount of data. However, if the data size is not large enough, some of the more "traditional" algorithm like Naive Bayes may perform better.

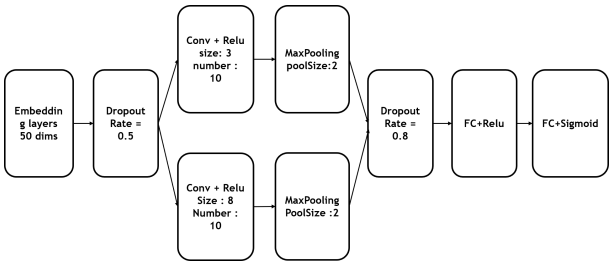

Figure 2: CNN architecture used for relevance classification.

Relevance Detection

The first functionality of the Biomimicry Explorer is to detect articles that are relevant to Biomimicry. Table 2 shows the prediction accuracy of the above classifiers based on the 10 % held out testing data. Here, CNN achieves the best performance with 77.88 percent accuracy. The architecture, shown in Figure 2, is a variant of the CNN architecture proposed by (Kim 2014), where we use two filter sizes with 3 and 8 instead of original 3 filter sizes, we used much fewer filters (10 instead of 100) based on experiment results and smaller embedding dimensions.

Function Classification

The next functionality of the Biomimicry Explorer is to organize the relevant article by their functions. We employ various multi-class classifiers to classify the input text into one of the eight pre-defined functions including *Protect from harm, Get resources, Breakdown, Process information, Maintain Community, Move or stay put, Modify and Make*, after it has found to relevant by the relevance classifier. Different machine learning algorithms are compared and some of the results are included in table 3 with Multinomial Naive Bayes achieving the best prediction accuracy. The CNN architecture was not competitive in this task and this can be attributed to he imbalance between classes and lack of sufficient data for several classes.

We perform post-hoc analysis of the best classifier to get important words from the input text which can be used as keywords to index the articles. Specifically, we use Local Interpretable Model-agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) to obtain class relevant keywords from the input articles. For a given class and sample,

model	mean	stdev
LogisticRegression	0.7294	0.0448
LogisticRegression (tf-idf)	0.7129	0.0413
RandomForest (BoW)	0.7388	0.0419
RandomForest (tf-idf)	0.7341	0.0562
naive Bayes (Bernouli)	0.7388	0.0516
naive Bayes (Bernouli tf-idf)	0.7388	0.0516
naive Bayes (Multinomial)	0.7365	0.0387
naive Bayes (Multinomial tf-idf)	0.7271	0.0451
SVM	0.7200	0.0443
SVM (tf-idf)	0.7153	0.0419
CNN	0.7788	0.0262

Table 2: Relevance classification results.

model	mean	stdev
LogisticRegression	0.4638	0.0277
LogisticRegression (tf-idf)	0.4553	0.0387
RandomForest (BoW)	0.4340	0.0512
RandomForest (tf-idf)	0.4511	0.0485
naive Bayes (Bernouli)	0.3447	0.0316
naive Bayes (Bernouli tf-idf)	0.3447	0.0316
naive Bayes (Multinomial)	0.4723	0.0530
naive Bayes (Multinomial tf-idf)	0.4681	0.0737
SVM	0.4596	0.0243
SVM (tf-idf)	0.4723	0.0485

Table 3: Function classification results.

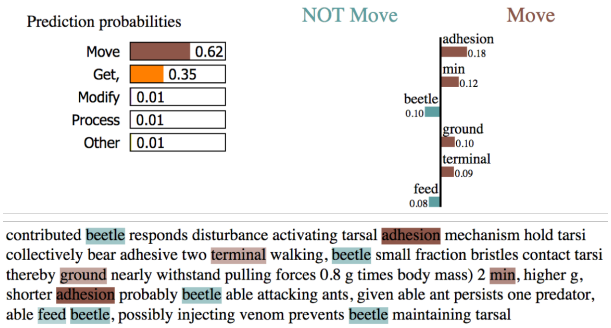

Figure 3: Visualizing word importances using LIME (Ribeiro, Singh, and Guestrin 2016) framework.

the LIME framework perturbs the sample in its neighborhood and learns a local interpretable linear classifier, whose weights give us the importance of words for the sample and class pair. Please see algorithm 1 defined by (Ribeiro, Singh, and Guestrin 2016) for the detailed algorithm. Figure 4 is an example of the output that generated by LIME on one of our abstract, the classifier predict the article belongs to class *Move or stay put*. We can see the predictor doing well on this example because it puts high weight on keyword like 'ground' and 'adhesion' that are understandable by human.

These important words are then used as keywords and to score sentences in text body. First, we calculate the weight of the sentences based on the average value of the phrases it contains Then, Apply a position weighting. Order each sentence from 0 to 1 equally based on the sentence number in the document. For example if there are ten sentences in a document, sentence nine's *position weighting* would be 0.9. This weight is then multiplied by the value calculated on step 1.

User Interface

The data outputs of this tool might be useful within the context of a more comprehensive application like AskNature, or via a stand-alone application. As a standalone service, users might explore cataloged records via a combination of keyword search and tag-oriented browsing, in which organism names and the eight pre-defined functions listed above would make up the primary tags. Users could use these mechanisms to identify articles describing potentially relevant biological analogies to their design challenges, and then follow external links to read the full text of the most compelling article matches. As they browse via tags, +/- flags would let users actively indicate how relevant and/or accurate particular classifications are. A new data point would be recorded for each positive or negative correlation indicated. In addition, whenever an external link was followed, a new data point could be recorded to indicate a positive correlation between the tag being filtered by and the active record, providing a passive method for users to improve the system over time.

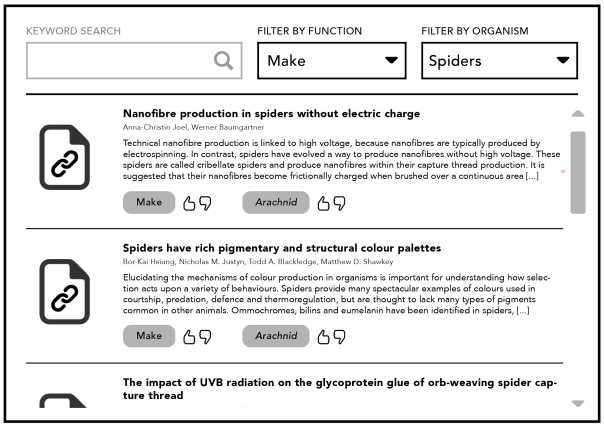

Figure 4: Potential (simiplified) interface for browsing cataloged data.

Discussion and Conclusions

In this paper, we present our approach to build a system that uses machine learning techniques to assess whether a scientific article could potentially serve as inspiration to a biomimicry invention. In addition, we classify relevant articles according to the first level in the Biomimicry taxonomy. One of the most challenging aspects of our work is collecting the appropriate data to train machine learning algorithms for the specified tasks. As such, we developed a crowdsourcing application based on serverless technologies which allowed us to collect data for scientific articles. Using the collected data, we devise several classifiers that show promising accuracies. Our system can help innovators identify articles describing relevant biological analogies to their design challenges and can also provide the users a system generated summarization highlighting relevant information. While this is only a first step towards automatic discovery of relevant biomimicry resources, it is a foundational step towards a scalable system that bridges the domains of biology and engineering to foster innovations inspired by nature.

References

Benyus, J. 1997. *Biomimicry: Innovation Inspired by Nature*. Harper Perennial.

Berger, A. L.; Pietra, V. J. D.; and Pietra, S. A. D. 1996. A maximum entropy approach to natural language processing. *Computational linguistics* 22(1):39–71.

Chakrabarti, A.; Sarkar, P.; Leelavathamma, B.; and Nataraju, B. 2005. A functional representation for aiding biomimetic and artificial inspiration of new ideas. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 19(2):113–132.

Cheong, H.; Li, W.; Cheung, A.; Nogueira, A.; and Iorio, F. 2017. Automated extraction of function knowledge from text. *Journal of Mechanical Design* 139(11):111407.

Dahl, D. W., and Moreau, P. 2002. The influence and value of analogical thinking during new product ideation. *Journal of Marketing Research* 39(1):47–60.

Deldin, J.-M., and Schuknecht, M. 2014. The asknature database: enabling solutions in biomimetic design. In *Biologically inspired design*. Springer. 17–27.

Gentner, D. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science* 7(2):155–170.

Goel, A. K. 1997. Design, analogy, and creativity. *IEEE expert* 12(3):62–70.

Hawkes, E. W.; Eason, E. V.; Christensen, D. L.; and Cutkosky, M. R. 2015. Human climbing with efficiently scaled gecko-inspired dry adhesives. *Journal of The Royal Society Interface* 12(102):20140675.

Helms, M.; Vattam, S. S.; and Goel, A. K. 2009. Biologically inspired design: Process and products. *Design Studies* 30:606–622.

Hooker, G., and Smith, E. 2016. Asknature and the biomimicry taxonomy. *INSIGHT* 19(1):46–49.

Hope, T.; Chan, J.; Kittur, A.; and Shahaf, D. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 235–243. ACM.

2017. Ibm bluemix openwhisk. `https://console.ng.bluemix.net/openwhisk/`.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* 137–142.

Johnson, R., and Zhang, T. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Jurafsky, D., and Martin, J. H. 2000. *Speech and Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall.

Kim, J.-D.; Ohta, T.; Pyysalo, S.; Kano, Y.; and Tsujii, J. 2009. Overview of bionlp'09 shared task on event extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, 1–9. Association for Computational Linguistics.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. *ArXiv e-prints*.

Kruiper, R.; Vincent, J. F.; Chen-Burger, J.; and Desmulliez, M. P. 2017. Towards identifying biological research articles in computer-aided biomimetics. In *Conference on Biomimetic and Biohybrid Systems*, 242–254. Springer.

LeCun, Y.; Bengio, Y.; et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361(10):1995.

Liaw, A.; Wiener, M.; et al. 2002. Classification and regression by randomforest. *R news* 2(3):18–22.

McCallum, A.; Nigam, K.; et al. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, 41–48. Madison, WI.

Moreno Grandas, D. P.; Blessing, L.; Yang, M.; and Wood, K. 2015. The potential of design-by-analogy methods to support product, service and product service systems idea generation. In *DS 80-5 Proceedings of the 20th International Conference on Engineering Design (ICED 15) Vol 5: Design Methods and Tools-Part 1, Milan, Italy, 27-30.07. 15*.

Nédellec, C.; Bossy, R.; Kim, J.-D.; Kim, J.-J.; Ohta, T.; Pyysalo, S.; and Zweigenbaum, P. 2013. Overview of bionlp shared task 2013. In *Proceedings of the BioNLP Shared Task 2013 Workshop*, 1–7. Association for Computational Linguistics Sofia, Bulgaria.

Pang, B.; Lee, L.; et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2):1–135.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Raibert, M.; Blankespoor, K.; Nelson, G.; and Playter, R. 2008. Bigdog, the rough-terrain quadruped robot. *IFAC Proceedings Volumes* 41(2):10822–10825.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: ACM.

Shu, L., and Cheong, H. 2014. A natural language approach to biomimetic design. In *Biologically Inspired Design*. Springer. 29–61.

Smith, C.; Bernett, A.; Hanson, E.; and Garvin, C. 2015. Tapping into nature.

Vandevenne, D.; Verhaegen, P.-A.; Dewulf, S.; and Duflou, J. R. 2016. Seabird: Scalable search for systematic biologically inspired design. *AI EDAM* 30(1):78–95.

Vattam, S.; Wiltgen, B.; Helms, M.; Goel, A. K.; and Yen, J. 2011. Dane: fostering creativity in and through biologically inspired design. In *Design Creativity 2010*. Springer. 115–122.

Vincent, J. F., and Mann, D. L. 2002. Systematic technology transfer from biology to engineering. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 360(1791):159–173.

Witten, I. H.; Frank, E.; Hall, M. A.; and Pal, C. J. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Icml*, volume 97, 412–420.

Yang, X.-S. 2014. *Nature-inspired optimization algorithms*. Elsevier.