**Question Answering with Encyclopedia of Life: Accessing Large-Scale Biological Knowledge**

Pramodith Ballapuram, Christopher Cassion, Abbinayaa Subramanian, Sungeun An, Robert Bates, Spencer Rugaber, Jennifer Hammock, Ashok Goel
Design & Intelligence Laboratory, School of Interactive Computing, Georgia Institute of Technology

**Abstract:** The practice of biologically inspired design requires access to general biological knowledge. In this paper, we describe AskEOL, a question-answering tool for accessing biological knowledge from Encyclopedia of Life (EOL), the world’s largest knowledgebase of biological taxa. AskEOL operates in the context of a virtual research assistant, Vera, that provides an interactive environment for building conceptual models of ecological systems and runs experimental simulations on those models.

**1 Introduction**

The practice of biologically inspired design (Benyus 1997; Vincent & Mann 2000) requires both broad knowledge of biology and deep understanding of biological systems (Goel, McAdams & Stone 2014). Without a broad knowledge of general biology, a designer may not know of biological phenomena relevant to a design problem, and without a deep understanding of relevant biological systems, a designer may not know what to transfer to the design problem and what not to, or how to abstract the relevant knowledge to make the transfer.

However, designers have either the breadth of knowledge of biological phenomena or the depth of understanding of biological systems needed for effective biologically inspired design. The question then becomes how can we provide designers with broad access to large-scale general biological knowledge as well as help them construct deep understanding of specific biological systems. In this paper, we address the first question of broad access to large-scale general biological knowledge. In related projects, we address the second question of deep understanding in the form of functional models of biological systems (Goel et al. 2012) and acquisition of the models from natural language documents (Rugaber et al. 2016).

One way of acquiring biological knowledge is through contextualized querying of a large general-purpose knowledge base, such as the Encyclopedia of Life (EOL) [http://eol.org/](http://eol.org/). In this paper, we describe AskEOL, a question-answering tool for accessing biological knowledge from EOL. AskEOL operates in the context of the Virtual Ecological Research Assistant (Vera) that provides an interactive environment for building conceptual models of ecological systems and running experimental simulations on those models. AskEOL uses IBM Watson’s Natural Language Classifier (Watson NLC; https://console.bluemix.net/#/store; Ferrucci et al. 2010; Lally et al. 2012) and Stanford’s CoreNLP tool ([https://stanfordnlp.github.io/CoreNLP/](https://stanfordnlp.github.io/CoreNLP/); Manning et al. 2014) tool to enable users to make natural language queries for the knowledge needed to build deep conceptual models of ecological phenomena.

**2 Encyclopedia of Life**

The Encyclopedia of Life (EOL) is the world’s largest knowledgebase of biological taxa. It presently contains 11 million records in 53 datasets (Parr et al. 2014a). Figure 1 shows a word cloud of the data in EOL. EOL’s TraitBank contains information about 340 traits of 1.7 million taxa

![Fig 1. EOL Data Word Cloud](image)

(Parr et al 2014b). Figure 2 shows part of a TraitBank webpage for a specific organism. The current version of AskEOL focuses mostly on the EOL’s structured knowledge indicated on the top right of the webpage shown in Figure 2.

**2.1 VERA**

Vera (An et al. 2018) is a conceptual modeling and agent-based simulation tool that can be used by citizen, student, and professional scientists alike. It allows its users to iteratively build ecological conceptual models and run experimental simulations based on those models. It does this using a custom made built-in compiler that enables the system to translate the conceptual models into simulations. Figure 3 illustrates Vera’s architecture that consists of MILA-S (Joyner & Goel 2014; Goel & Joyner 2015; Agarwal et al. 2018), an interactive tool for conceptual and simulation modeling, as the frontend and EOL as the backend.

To accurately model and represent a desired ecosystem, the user may need information regarding the different taxa represented in the system and their properties such as their habitats and lifespans. Citizen scientists may not be aware of these values and thus require access to this information.
Although web-based search engines can access a lot of information, it can be hard for a novice user to filter the needed information from the millions of available sources.

2.2 Watson

IBM Watson’s Bluemix services provide tools for Natural Language Processing that can be integrated with any application. AskEOL uses Bluemix to understand the queries users make of Vera. We experimented with various tools offered by the Bluemix services and found out the Natural Language Classifier tool (Watson NLC) was the most useful for our use case.

3 Methodology

AskEOL is a domain-specific Question Answering tool geared towards ecology. In this section, we explore the technical components that make up AskEOL. First, we discuss the question templates used to create the dataset to train the Watson NLC, next we go into details on how the dataset was created. Then, we explain the two main modules of AskEOL: the question classifier and the target taxon extraction component. Finally, we talk about how the answers are retrieved.
3.1 Question templates

AskEOL can answer a subset of the types of questions that a user of Vera may need to answer in order to construct a model. As of now, these question types are:

(i) life span: how long a species lives
(ii) habitat: where an animal lives
(iii) rate of reproduction: how often does an animal reproduce
(iv) diet: what an animal eats

These question types refer to user tunable properties of the conceptual model simulated by Vera. The answer returned by AskEOL can then be used as parameter values to produce more accurate simulations. Some example templates can be seen in Table 1.

<table>
<thead>
<tr>
<th>Question Types</th>
<th>Example 1</th>
<th>Example 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Life Span</td>
<td>What is the life expectancy of x?</td>
<td>What is the longevity of x?</td>
</tr>
<tr>
<td>Habitat</td>
<td>What climate do x prefer?</td>
<td>What is the natural habitat for x?</td>
</tr>
<tr>
<td>Rate of Reproduction</td>
<td>What is the level of fertility of x?</td>
<td>What is the primary level of fertility in x?</td>
</tr>
<tr>
<td>Diet</td>
<td>Does x feed on y?</td>
<td>What does x eat?</td>
</tr>
</tbody>
</table>

Table 1. Sample question templates

3.2 Creating the dataset

The dataset for training the Watson NLC tool was created using the four question types above as constraints. We manually came up with several unique templates for ways in which a question could be asked. Instead of specifying a specific taxon for the templates, we used wildcard placeholders. This helped creating the question much faster. A total of 360 training questions were generated in order to use it for training, the placeholders needed to be replaced by real taxon names. This is to ensure that the algorithm can determine the proper noun used to name a taxon as a possible feature. To that end, a list of taxa names was obtained from EOL’s Global Biotic Interactions database (Poelen et al. 2014) and a script was used to fill the placeholders with randomly chosen names. Since the authors were familiar with the training set and biases of the system, we crowdsourced a test set from

<table>
<thead>
<tr>
<th>Sample Question</th>
<th>Class</th>
<th>Answer from AskEOL</th>
</tr>
</thead>
<tbody>
<tr>
<td>How long do armadillo live?</td>
<td>lifespan</td>
<td>Armadillo lives for 180 months.</td>
</tr>
<tr>
<td>Where can I find deer?</td>
<td>habitat</td>
<td>Dense forest and planted areas.</td>
</tr>
<tr>
<td>What do dolphins eat?</td>
<td>diet</td>
<td>Dolphin eats Fish, Crab, Squid.</td>
</tr>
<tr>
<td>How old can tiger be?</td>
<td>lifespan</td>
<td>Tiger lives for 50 years.</td>
</tr>
<tr>
<td>How often do tigers reproduce?</td>
<td>Rate reproduction</td>
<td>Tiger gives birth at an average of once every 730 days.</td>
</tr>
<tr>
<td>At what frequency do sharks reproduce</td>
<td>Rate reproduction</td>
<td>Sharks gives birth at an average of once every 730 days.</td>
</tr>
<tr>
<td>What kind of terrain do squirrels live in?</td>
<td>habitat</td>
<td>Squirrels can be found in soil, prairie.</td>
</tr>
<tr>
<td>How often do bats get fertile?</td>
<td>Rate reproduction</td>
<td>Bat gives birth at an average of once every 365 days.</td>
</tr>
<tr>
<td>Do deer eat dolphins?</td>
<td>diet</td>
<td>Deer eats Acorns, Fruit, Grass</td>
</tr>
<tr>
<td>How old is a cow when it dies?</td>
<td>lifespan</td>
<td>Cow lives for 12 - 20 years.</td>
</tr>
</tbody>
</table>

Table 2. Sample responses from AskEOL.
individuals who were not involved in the process of creating the training set. They were simply provided the four question types and one sample question for each class. In this way, a test set of 120 questions was created.

3.3 Question classification module
AskEOL’s question classification module is responsible for classifying questions asked by the user into one of the four question types above. Initially, we trained a model to identify the class of the question and target taxon using an entity-relationship extraction technique. However, this did not provide us with satisfiable results. Thus, created one module for classifying the question and a second module to identify the target taxon.

3.4 Finding the target species of a question
Once AskEOL identifies and classifies the type of question being ask, the next step in the process is to understand who or what the query is about (i.e. target taxon). This module, which primarily uses the Core NLP tool, parses the user’s query and returns the target taxon. The target taxon of the query is the taxon the user wants to obtain information about. For example, in the question "What does a dolphin eat?", the target taxon is dolphin. If the query is, "How often do sea lions mate?", the target taxon is sea lions. The taxa matching is currently done by the EOL pipeline.

Like the example questions, there are several cases when identifying target species. The first case is if the name of the taxon is a singular noun (i.e. dolphin), and the second is if the taxon name is multi-part (i.e. hammerhead shark), the third is if we have more than one taxon in the question such as "Do dolphins eat sharks?". Each of these is handled differently. For questions with a simple taxon name as in the first case, we make use of Parts of Speech tagging to identify proper nouns.

To handle the latter cases, we use Stanford’s CoreNLP tool for dependency parsing. In the example of hammerhead shark, we use dependency parsing to identify instances when the words such as hammerhead and shark must be combined to form a single species. A set of rules were created to handle these cases. The rules primarily revolved around identifying the "noun subject" that is the root of the dependency graph. An example rule is if more than one noun is present, we create a Dependency Parse Tree and return the subject noun along with any compounding modifiers as the noun subject.

If these techniques do not yield a proper taxon name, i.e. no word is classified as the taxon name, we execute a catch all procedure. This part of the module contains a balanced tree populated with taxa names known to the system. To make use of it, we query the tree and attempt to perform a word match between the words in the query and the taxa names present in the tree.

3.5 Answer retrieval module
Now that AskEOL understands the question, the next step is to answer it. This module returns an answer to the user’s query based on the provided input. The input in this case is the question type classification and target taxon extracted by the previous modules. AskEOL’s primary information database is the Encyclopedia of Life (EOL). If the information cannot be extracted from EOL, then we make use of another data source called A-Z animal (https://a-z-animals.com). Table 2 shows a sample of responses from AskEOL.

3.5.1 Validating the input
The first step of this module is to check the information passed in as input for correctness. First, this module checks that the question class returned matches one of the four supported types. Second, the module checks if the target taxon is a valid input (i.e. not null, or empty). If any of these checks fail AskEOL returns "Sorry AskEOL can’t answer this question." to the user.

3.5.2 Extracting the answer
If both these checks pass, then AskEOL starts interacting with EOL through its Application Programming Interface (API). Due to the way that EOL is structured, information regarding a taxon cannot be accessed directly using the taxon’s name. AskEOL must first obtain the EOL unique identifier (id) associated with the taxon. Once that unique id is obtained, it is used to access the right location in EOL’s databank where information about the species is stored. This information is then returned to AskEOL for local parsing. The information returned contains all attributes and information that EOL has regarding the particular taxon. This includes both structured and unstructured data. We parse through the structured data since the queries AskEOL support are limited.

The parsing is done through a mapping between the class of the question and the key’s in the databank. For example, the AskEOL class lifespan is mapped to the key “total life span” in the information returned by the EOL databank. If neither the taxon nor the key is present in the EOL databank, we query the A-Z-animals databank (https://a-z-animals.com/) similarly. If AskEOL still cannot find the answer, it outputs a message announcing that "AskEOL can’t find the answer." Figure 4 shows the steps in processing used by AskEOL.

3.6 Illustrative Example
Let us suppose that the question asked by the user is “How long do armadillo live?” IBM Watson’s NLC identifies the class to be lifespan. Stanford Core NLP tool identifies the species to be armadillo. AskEOL maps the name armadillo to it’s unique ID in EOL TraitBank; the ID serves as an index into the TraitBank data for the
species. AskEOL now map “lifespan” to the corresponding key “total life span”. It extracts the data corresponding to this key in the EOL TraitBank, and frames the data as a sentence to the user.

4 Results

To assess the success of the AskEOL tool we performed empirical experiments as well as user studies. The empirical experiments were to quantify the accuracy of our classifier and the accuracy at which target taxa in queries are identified. The user study was conducted to assess the usability of the tool and to confirm if the tool enhanced the experience of users of Vera.

4.1 Preliminary experiments

The trained Watson NLC obtained an accuracy (True Positives/ Total size of Test set) of 84 percent on the crowd sourced test set of queries. AskEOL successfully identified the target taxon with an accuracy of 60 percent. A few sample questions and the responses obtained from AskEOL can be seen in Table 2.

4.2 Design of user study

A small user study with a group of 3 students was conducted. The study focused on the usability of the tool rather than the accuracy of answers. We provided the students with a set of sample queries and the supported categories of questions, along with a list of target taxa to choose from. The list of target taxa was provided to ensure that the databanks we provided had the information needed to answer a query asked by a user. The participants were asked to use the AskEOL tool first and then use a search engine of their choice to obtain the answer for the same set of queries. The participants were then asked to answer a questionnaire on their experience of using the tool. Most questions had answers from a scale from 1 to 4, with 4 being the most positive and 1 the most negative response in terms of how good/useful the tool was for the task.

All participants found the tool “Easy” to use (all gave a score of 3 on a scale of 4). Two of them found AskEOL “Useful” (3 on a scale of 4) in the context of conceptual modeling in Vera, with the third student unsure about its usefulness. In terms of comparison with the Google search engine, one student preferred the search engine, while the other two scored Google and AskEOL equally.

5 Discussion

This work builds on our earlier work on supporting biological inspired design by accessing biological knowledge from natural language documents (Goel 2016). In the previous work, we used IBM’s Watson system for answering questions about biological phenomena based on a corpus of biology articles. A major limitation of the previous work was that we had to manually annotate the articles. In practice, this meant that we were limited to a small corpus.

In this paper, we have build a novel question-answering system called AskEOL to support ecological modeling by leveraging the Encyclopedia of Life, a large knowledgebase with built-in annotations. By giving the user the ability to accurately model their desired ecosystems, they can learn more about the phenomena they are attempting to model. This enables them to properly tune the parameter values the simulations and gain more meaningful insights from the simulation results.

A limitation of the current work is that AskEOL presently works only with EOL’s structured data. However, a large portion of EOL contains unstructured data. Future work includes leveraging this unstructured data to answer users’ natural language queries. Further, to support and attract a wider range of citizen scientists, we would like to increase the classes of questions that AskEOL can answer including categories such as body mass, preyed upon, providing paternal care etc. Additional future work includes improving AskEOL’s capability to identify target taxa, leverage state-of-the-art methods and applications such as training our own named entity recognizer exclusively for animal taxa. Based on the feedback from the user study, we also plan to improve the format in which a response from AskEOL is presented.
Acknowledgements: This research is supported by an US NSF grant #1636848. We are grateful to IBM for providing access to their Bluemix services. We also thank David Joyner and Animesh Mehta for several discussions.

References