# Explanation Container in Case-Based Biomedical Question-Answering

Prateek Goel[1], Adam J. Johs[1], Manil Shrestha[2], and Rosina O. Weber[1]

Dept. of Information Science, Drexel University, PHL 19104, USA [1]
Dept. of Computer Science, Drexel University, PHL 19104, USA [2]

**Abstract.** We present the design of the Explanatory Agent, a case-based agent conceived to answer biomedical queries. The Explanatory Agent is an autonomous relay agent within the multi-agent architecture of the Biomedical Data Translator, an initiative by the National Center for Advancing Translational Sciences. To answer queries, the Explanatory Agent seeks knowledge from multiple sources, ranks the results derived from such knowledge, and explains the ranking of results. The design of the Explanatory Agent encompasses five knowledge containers–the four original knowledge containers and one additional container for explanations, the Explanation Container. The design of the Explanation Container is case-based and encompasses three knowledge sub-containers. We utilize a drug-repurposing use case to illustrate the Explanatory Agent's capacity for answering biomedical queries.

**Keywords:** Case-based reasoning, knowledge containers, explanation container, translational research, question-answering

## 1 Introduction

In this paper, we present the design of the Explanatory Agent, a case-based automated relay agent that answers biomedical queries by accessing multiple knowledge providers, ranks results, and explains their ranking. The Explanatory Agent (xARA) is part of and is sponsored by the National Center for Advancing Translational Sciences (NCATS) Biomedical Data Translator [1] (Translator).

The Translator is an NCATS's endeavor motivated by the breadth, complexity, and heterogeneity of biomedical knowledge and data that remain challenging hurdles for researchers in translational research [1, 2, 3]. The ability to submit a biomedical question to a centralized system to quickly receive actionable information is an attractive goal, but has been complicated due to numerous factors. These include the use of competing or otherwise separate data integration services, where each may provide partial answers to a question, but also require some level of expertise to use and integrate for interpretable answers. In consequence, NCATS's Translator [1] aims to link existing biomedical knowledge sources and data integration services to deliver actionable answers to crucial questions in support of patients' well-being [3]. Translator includes six autonomous relay agents; the Explanatory Agent is one of them.

The NCATS defines translational research as "*. . . the endeavour to traverse a particular step of the translation process for a particular target or disease.*" p. 455 [4] While "*. . .translation is the process of turning observations in the laboratory, clinic and community into interventions that improve the health of individuals and the public — from diagnostics and therapeutics to medical procedures and behavioural changes.*" p. 455 [4]. The term that inspired the name of the system discussed herein is widely used in biomedical science.

Translator focuses on mitigating two real-world issues associated with biomedical data. The first problem stems from the diversity of biomedical data. Biomedical data are spread in different areas of research without standardization, connection, or a common vocabulary [3]. The second problem refers to the vast scale of biomedical data, which may be too complex for humans to comprehend without elaborate methods. For example, there were 1,613 molecular biology databases available as of January 2019 [5]. For these reasons, the Translator aims to link data from such databases with patient data, chemical and pharmaceutical data, clinical trials, biomedical ontologies, and possible augmentations that can support scientists while designing and refining the research questions they should prioritize. The ultimate goal of the Translator is to streamline the path from a biomedical research lab to the patient's bedside [1, 2, 3]. Along this path, the Translator shall accelerate translational research, help generate new hypotheses, and drive new innovations in clinical care and drug discovery [3].

The Translator system design incorporates three sets of interlinked components, namely, the Autonomous Relay System (ARS), the ARAs and Knowledge providers (KPs) [1]. KPs are Translator's internal aggregators, each specializing in one or many types of knowledge. ARAs are Translator's reasoning agents, tasked with selecting which KPs to ask for data, organizing results, ranking results, and preparing responses. The ARS breaks down user queries into the internal knowledge graph representation language and transmits them to ARAs, such as xARA, to provide actionable responses via relations obtained using the KPs. To realize this multi-agent system, the Translator's agents utilize a standardized API language created within the consortium and the Biolink model[1] as a data model.

xARA is a case-based ARA that uses five knowledge containers to rank and explain results received from KPs to answer user queries transmitted by the ARS. xARA distinguishes itself from other ARAs by providing ranked results to biomedical queries based on explanations. To execute received queries, xARA uses the case-based reasoning (CBR) methodology to make decisions with respect to which biomedical knowledge sources to refer to and obtain results from. This paper describes the design of xARA with the four original knowledge containers [6, 7] along with an additional container for explanations. The Explanation container, among others, can produce explanations to support and provide evidence for biomedical relations. To realize its full potential, the Explanation container's design is case-based and maintains its own set of knowledge containers.

[1] https://biolink.github.io/biolink-model

Section 2 provides a discussion of the knowledge containers model [6, 7], which are important background to this paper. The contribution of this paper is described in Section 3 with the details of the xARA design.

## 2 Background on CBR Knowledge Containers

CBR is a methodology to solve new problems by reusing previously stored problem-solution pairs. The CBR methodology can be described from two modeling perspectives, namely, the CBR Cycle [8] and the Knowledge Containers model [6, 7]. The CBR cycle presents the CBR methodology as a sequence of steps to execute the steps *retrieve*, *reuse*, *revise*, and *retain*. The knowledge containers model observes the CBR methodology from a knowledge-based perspective where its knowledge is distributed across different modules.

The four original knowledge containers are Vocabulary container, Similarity container, Case Base container and Solution Transformation container. xARA implements two of the steps in the CBR cycle, *retrieve* and *reuse*, and is designed around the knowledge containers model.

The Vocabulary container enables the agent to recognize entities from the domain. It helps to complement incomplete input by using vocabulary knowledge. The knowledge in the vocabulary container is mostly declarative (e.g., ontological), and its processes bridge the gap between varying levels of specificity of incoming problems and stored cases [7].

The Case Base container retains the contextualized experiences represented in cases, providing coverage for problems the agent has to solve. Cases are often represented as problem-solution pairs, and an outcome may be used to keep track of how successful a solution is when applied to a problem. Cases can be based on real-world experiences, artificially crafted, or adjusted from existing data.

The Similarity container keeps the knowledge required to assess similarity between a new case and the cases in the case base. For this, the Similarity container accesses the Case Base container. The Similarity container may access the Vocabulary container when cases are at different levels of specificity, or their ontological structure is relevant for assessing similarity. It is typical in CBR that the similarity functions assess similarity at the local level and then aggregate them into a global score, which is used to represent whether a candidate case comprises enough similar attributes to solve for the new case.

The Solution Transformation container consists of the knowledge required to transform the reused case or cases to correctly solve the new case. Adaptation is an important component in this container. Adaptation may be needed when there are significant differences between the new case and the reused case(s). This container may include substantial knowledge bases, and can be case-based.

## 3 The Explanatory Agent Design

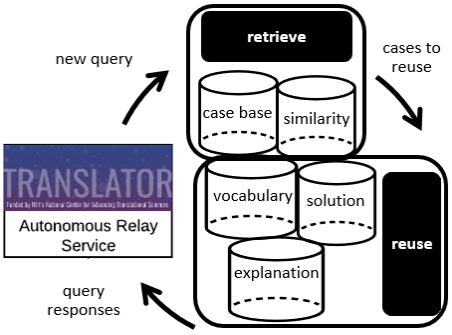


**Fig. 1.** xARA implements *retrieve* and *reuse* with five knowledge containers: Case Base, Similarity, Vocabulary, Solution, and Explanation

The xARA is one of the Translator's ARAs that obtains, executes, and provides ranked and explained responses for user queries. The xARA uses biomedical expertise embedded in its cases to decide which KPs within the Translator to consult to produce answers to user's queries. Aiming to provide useful answers, xARA relies on explanation methods, natural language understanding models, or simple rules, to rank results and indicate to users the reason a result is ranked.

The xARA implements two steps of the CBR cycle, *retrieve* and *reuse*, using five knowledge containers (Fig. 1), namely, Case Base, Similarity, Solution, and Vocabulary, with an additional container for explanations. The Explanation container is case-based and uses sub-containers in its design. In Fig. 1, the steps *retrieve* and *reuse* are grouped with the containers they use. Note in Fig. 1 that adjacent knowledge containers are those that are connected. The Similarity container needs to access the Case Base container and potentially the Vocabulary container. The Solution container may also use the Vocabulary container and it is connected to the Explanation container. The description of the containers in xARA follows.

**The Case Base Container** The primitive cases in xARA are based on queries and responses that the KPs can return. Additionally, derived cases are created based on expansions of primitive cases. The indexing of cases includes node categories and predicates. Node categories may represent biomedical concepts such as diseases, drugs, genes, chemical reactions, and proteins. Predicates represent semantic associations between those concepts, such as *treats*, *involved in*, or *contributes to*. The Case Base container is used when the *retrieve* step uses the Similarity container to find cases.

**The Similarity Container** The Similarity container computes scores for candidate cases to determine which ones to reuse. The xARA relies on a threshold to determine which cases will be reused. Multiple cases may be reused. The Similarity container relies on the Vocabulary container to assess similarity between nodes and predicates. The Biolink ontology [2] is used to determine how close two predicates are by way of structured similarities. There is also an associated level

[2] https://biolink.github.io/biolink-model

of specificity, given that when a specific predicate is not available, a more general predicate can be used. For example, when a query asks for *drugs that inhibit a certain protein*, if this exact relation is not available, the more abstract relation *drugs that are correlated with* that same protein may be used. The global similarity aggregates local functions and considers the relative relevance of each of the elements considered such as nodes and predicates. Their combined functionality identifies the most relevant cases for *reuse*.

**The Vocabulary Container** The Vocabulary container in xARA provides support for similarity functions as mentioned above, and processes the new case into the representation used in similarity assessment. The queries may or may not be represented in the scope of the agent's knowledge or be limited in providing enough detail about the query, so the Vocabulary container helps the agent convert the query into the format the Similarity container needs for performing retrieval and identifying cases for *reuse*. For example, queries may indicate the specific name of a disease rather than the category *Disease*. The cases are indexed at the level of categories and not by specific names of entities. The functions to assess the category of a named entity is in the Vocabulary container, which enables the new case to be represented at the same specificity as the candidate cases.

**The Solution Container** The transformation of reused cases to produce solutions include multiple steps in xARA. The first step is to extract the strategy adopted in the reused case. Because Translator data are represented as knowledge graphs, queries may consist of different configurations of nodes and edges with various hops. The case problems however are all one-hop triplets, so there may be transformations and combinations of previous cases. The Solution container also houses the functions to produce an output in accordance to Translator standards. Before preparing the final response, results received from KPs are explained and ranked in the Explanation container.

**The Explanation Container** The need for a container for explanation data and functions stems from the fact that xARA uses various methods and external data sources to produce explanations to the results obtained from multiple KPs. The explanation container could have been potentially designed as a standalone tool, but it has not; the explanation is a functionality offered by the xARA to justify how it ranks results obtained by other agents.

The explanation container relies on its own set of sub-containers: xCase Base, xSimilarity, and xSolution. As mentioned earlier, xARA explains results obtained from KPs. Consequently, the explanation functions are called after KPs return results. The results and the KP that provided the results are the input to the explanation methods, and thus describe the problems to be solved by explanation cases – xCases. xCases are retained in the xCase Base. The solution for xCases are an explanation approach. The explanation approach may utilize an

explainable artificial intelligence (XAI) approach (*e.g.*, [9]), it may simply cluster results, utilize a rule based on the methods adopted by the KP, or use methods based on fine-tuned language models such as BioBert [10].

The xSimilarity container is required because the knowledge that makes a result similar to another, such that their explanations can be interchangeable, differs from the knowledge used for cases in the Case Base container described above. To distinguish them, the cases in the main Case Base container are referred to as query cases, while those in the Explanation container, we refer to as xCases. The similarity functions to compare nodes and predicates may be accessed from the main similarity container, but they would be aggregated differently in the explanation container, thus requiring this xSimilarity container.

The xSolution container executes the explanatory functions. However, the Explanation container does not require a designated vocabulary container because it does not need a different vocabulary as the domain is the same and the vocabulary is the same throughout the agent.

## 4 Use Case

To demonstrate xARA's ability to query KPs using CBR and to provide contextual explanations for those results, we collaborated with other members of the Translator consortium to work on several exploratory use cases. One such use case would be drug re-purposing for Kennedy Disease (KD).

KD is considered a rare disease, meaning that it fits the criteria of affecting fewer than 200,000 people in the US annually. It is also monogenic, meaning that it arises from the mutation or dysfunction of a single gene, in this case the androgen receptor gene *AR*. Clinically, KD presents itself as a progressive neuromuscular degenerative disorder almost exclusively affecting males. The age of onset occurs between the ages of 20 and 50 with symptoms such as muscle weakness, cramps, and difficulty speaking and swallowing. Other symptoms include facial weakness, numbness, infertility, tremors, and enlarged breasts.

The *AR* receptor is responsible for transferring signals from male sex hormones, like testosterone, across cellular membranes to affect function in a number of cell types. It is unknown exactly how mutations in AR contribute to the KD phenotype, but it is understood that females are protected from the disease, even if they carry the mutation, since they have much lower levels of circulating testosterone. [3]

For this use case, the query logic was straightforward and exploratory in nature. Two queries were used in this case. The first step was to identify genes associated with KD. As expected, only *AR* was returned. Since this was our only result, we fed it into the second query which identified drugs that target the *AR* gene (NCBI:367). Of the 194 returned, the result associating *hydroxyflutamide* with *AR* was submitted for explanation.

Similar to previous results, *hydroxyflutamide* was identified as an *antagonist* of *AR*. This is unsurprising as antagonism, downregulation, and inhibition are

[3] https://rarediseases.org/rare-diseases/kennedy-disease/

similar and common general mechanisms of drug-like agents. In this case, a direct inhibition of *AR* by *hydroxyflutamide* is logically consistent with what we know of KD, although potentially impractical. We understand that females are protected from KD by low circulating testosterone, therefore it follows that inhibiting testosterone from interacting with *AR* would attenuate KD phenotypes. However, the long-term effects of using male hormone blockers in KD patients would need to be carefully considered by clinicians. More critically, off-target effects of drugs like *hydroxyflutamide* would also need to be carefully examined before considering further investigation. In fact, one potential off-target effect is mentioned within the same extracted passage in this example: inhibition of IL-6.

### 4.1 Use Case Considerations

Although the caveats mentioned in the previous section are important considerations, we maintain that decisions regarding drug re-purposing or any subsequent scientific research endeavor is beyond the scope of xARA and that of Translator as a whole. Our goal is to provide supportive, contextual, and semantically-rich explanations to scientific assertions returned by KPs as well as a mechanism to relay user queries to the most relevant KP. This is in line with Translator's goal of integrating multiple heterogeneous data and knowledge sources toward providing insights into the relationship between molecular and cellular processes and the signs and symptoms of diseases. In short, our agent is meant to augment, not replace, the workflow of biomedical and translational researchers.

Finally, it is important to understand that the use case presented above comes with the limitation that we selected individual answers for simplicity. In the full, standard use of xARA, all results would be expanded in subsequent batch queries to provide many more results than were demonstrated here. It is at this point that the ranking feature becomes crucial, as it will serve to filter and prioritize responses for user exploration.

## 5 Conclusion and Future Work

In this paper, we introduced the design of xARA, one of the agents in the Biomedical Data Translator, as an application of CBR in healthcare. The presentation of xARA was focused on its design that relies on the Knowledge Containers model proposed by Professor Michael Richter[6, 7].

One important observation from the experience with xARA was the realization that a vocabulary sub-container was not needed for the explanation container. We believe this may be true often as the vocabulary of the system reflects a user context. Consequently, once the user context is the same, the vocabulary would not change regardless of how many containers or CBR steps are case-based.

One limitation of this presentation is lack of related works. This paper is an illustration of using CBR in the health sciences but does not aim to contribute to the biomedical question-answering approaches and therefore analyzing those

works would be out of scope. The main contribution is the use of a container for explanations. With respect to its design, an ideal evaluation would be to compare it against alternative designs. The verification of the design will be done in the context of the Translator consortium. For validation, we will conduct user studies given the subjectivity of explanations.

**Acknowledgments** The authors would like to thank Professor Richter (*in memoriam*) for his many lessons. Thanks to J. Gormley, T. Zisk, and D. Corkill for their collaboration and feedback, and E. Hinderer III for his help with the use case. Support for the preparation of this paper was provided by NCATS, through the Biomedical Data Translator program (NIH awards 3OT2TR003448-01S1).

## References

[1] The Biomedical Data Translator Consortium. "Toward A Universal Biomedical Data Translator". In: *Clinical and Translational Science* 12.2 (2019), pp. 86–90. DOI: https://doi.org/10.1111/cts.12591. eprint: https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12591. URL: https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/cts.12591.

[2] Christopher P. Austin, Christine M. Colvis, and Noel T. Southall. "Deconstructing the Translational Tower of Babel". In: *Clinical and Translational Science* 12.2 (2019), pp. 85–85. DOI: https://doi.org/10.1111/cts.12595. eprint: https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1111/cts.12595. URL: https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1111/cts.12595.

[3] Stanley Ahalt et al. "Clinical Data: Sources and Types, Regulatory Constraints, Applications". In: *Clinical and Translational Science* 12 (2019), pp. 329–333.

[4] Christopher Austin. "Translating translation". In: *Nature Reviews Drug Discovery* 17 (Apr. 2018). DOI: 10.1038/nrd.2018.27.

[5] Daniel J Rigden and Xosé M Fernández. "The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection". In: *Nucleic acids research* 47.D1 (Jan. 2019), pp. D1–D7. ISSN: 0305-1048. DOI: 10.1093/nar/gky1267.

[6] Michael M. Richter. *The Knowledge Contained in Similarity Measures*. Invited Talk at the First International Conference on Case-Based Reasoning, ICCBR'95, Sesimbra, Portugal. . 1995.

[7] Michael M. Richter and Rosina O. Weber. *Case-Based Reasoning: A Textbook*. Springer Publishing Company, Incorporated, 2013. ISBN: 364240166X.

[8] Agnar Aamodt and Enric Plaza. "Case-based reasoning: Foundational issues, methodological variations, and system approaches". In: *AI communications* 7.1 (1994), pp. 39–59.

[9] David Gunning and David W Aha. "DARPA's explainable artificial intelligence program". In: *AI Magazine* 40.2 (2019), pp. 44–58.

[10] Jinhyuk Lee et al. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining”. In: *Bioinformatics* 36.4 (2020), pp. 1234–1240.