Irrelevant Explanations: a logical formalization and a case study

Simona Colucci[1,*], Tommaso Di Noia[1], Francesco M. Donini[2,*], Claudio Pomo[1] and Eugenio Di Sciascio[1]

[1]Politecnico di Bari, Via Orabona 4, 70125, Bari, Italy
[2]Università della Tuscia, Via S. Maria in Gradi, 4, 01100 Viterbo, Italy

Abstract

Explaining the behavior of AI-based tools, whose results may be unexpected even to experts, has become a major request from society and a major concern of AI practitioners and theoreticians. In this position paper we raise two points: (1) *irrelevance* is more amenable to a logical formalization than relevance; (2) since effective explanations must take into account both the context and the receiver of the explanations (called the explainee) so it should be also for the definition of irrelevance. We propose a general, logical framework characterizing context-aware and receiver-aware irrelevance, and provide a case study on an existing tool, based on Semantic Web, that prunes irrelevant parts of an explanation.

1. Motivation

Explanations services of AI tools are likely to provide one of the main interaction modalities between human users and AI-powered assistive technologies. Such an explaining modality may be useful also for AI experts, when the AI tool results surprise its very designers [1]. Given the raising importance of explanations, scholarly literature now abounds of studies about several types of explanation services, in various application scenarios.

Usually the explanation is considered as an understandable description of the results obtained. Yet, any explanation act involves a trade-off between relevant and complete explanations for whom the explanation is given to—what we will call the *explainee* in this paper. Generally speaking, demonstrating the relevance of knowledge is rather hard: the feeling about returned information is, in fact, fully subjective and what is interesting for a user may be completely useless for another one [2]. As an example, in the Berkeley Deep Drive-X (eXplanation) Dataset[1] [3] about self-driving car systems, the reason why the car is proceeding on a lane may just be that *(i) "there is nothing on its lane"*. While truthful, some user may (subjectively) find obvious that the car proceeds, given the user's knowledge that *(ii)* a destination was set for the car; *(iii)* in the absence of obstacles, the car is supposed to proceed to the destination, and that *(i)*

XAI.it 2023 - Italian Workshop on Explainable Artificial Intelligence
*Corresponding author.
simona.colucci@poliba.it (S. Colucci); tommaso.dinoia@poliba.it (T. Di Noia); donini@unitus.it (Francesco M. Donini); claudio.pomo@poliba.it (C. Pomo); eugenio.disciascio@poliba.it (E. Di Sciascio)
0000-0002-8774-4816 (S. Colucci); 0000-0002-0939-5462 (T. Di Noia); 0000-0003-0284-9625 (Francesco M. Donini); 0000-0001-5206-3909 (C. Pomo); 0000-0002-5484-9945 (E. Di Sciascio)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)
[1]https://github.com/JinkyuKimUCB/BDD-X-dataset

follows from *(ii)+(iii)*. Observe that a complete explanation may involve all three statements, leading to what would be perceived as a redundant explanation—that is, an explanation full of details that may be considered truthful but irrelevant, since they are already known by the explainee. Observe that a simply redundant explanation contains both details that are known to the explainee, and details that are not known, while in a completely irrelevant explanation all details were already known.

In this position paper, we want to establish some objective criteria for defining knowledge surely *irrelevant*, by elaborating on (and generalizing) ideas that we presented in a more restricted context [4]. In general, we observe that humans are usually not interested in being explained:

1. information that is true also for situations different from the one being explained, and in particular, information that is always true in general;
2. information they already know.

However a logic formalization of these widespread ideas is still missing, leading to direct implementations that, although justified by the above intuitions, are tailored to the specific application.

Stemming from the criteria above, we attempt a logical formalization of *irrelevant explanation*, in Section 2. In Section 3, we show the benefits of adopting such a formalization when explaining the similarity of groups of RDF resources. A final section concludes the paper.

2. Formalizing Relevant Explanations

To formalize irrelevance, we lay down several hypotheses about how to logically represent the setting in which an explanation arises. We are aware that some hypotheses may be questionable, but we consider them all necessary for a logical formalization of a relevant/irrelevant explanation. The elements we give a name to are: a deterministic system S, an explainee e, an input I to S. On input I, S outputs a result R, which e asks to explain. We suppose that:

1. both the input and the output (or, their descriptions) can be expressed as formulas I, R
2. the characteristics of S can be represented by a logical theory T_S such that $T_S \cup \{I\} \models R$; observe that if S were not deterministic, a more complex statement involving probabilities would be necessary
3. the knowledge possessed by e—*i.e.*, information e consciously knows—can be represented as another logical theory T_e
4. an explanation is formed by n sentences (*chunks*), each sentence stating the truth of a logic formula $E_i, i = 1, \ldots, n$, so that the entire explanation E can be represented as a conjunction $E \doteq E_1 \wedge \cdots \wedge E_n$
5. an explanation—although possibly irrelevant—is always truthful with respect to the particular behavior of system it explains, that is, $T_S \cup \{I\} \models E$.

We now motivate and discuss the above assumptions.

Assumption 1 seems rather straightforward: it is always possible to represent the inputs and the results of a system in some formal language.

Assumption 2 may seem too strong for numerical, nonlinear AI systems; however, it does not pretend to *completely* describe the inner functioning of S; only the fact that inputs and outputs can be logically related by T_S.

Assumption 3 takes into account both general knowledge and specific knowledge that can be attributed to the explainee. For example, for a physician an ontology of medical knowledge—*e.g.*, "Antibiotics cure bacterial infections"—can be added to general knowledge about the world.

Assumption 4 is necessary when an explanation is a complex argument, expressed as several sentences. The correspondence between sentences and formulas will be necessary for Point 2 below.

Finally, **Assumption** 5 is just a logical way to express a natural requirement: explanations should always be truthful with respect to how S works on input I. For example, if the counterfactual explanation given by S for denying a loan was "If the monthly income raises by 25%, the loan could be granted", one expects that just raising the monthly income (changing nothing else) the loan would *really* be granted. In formulas, if to explain the result R on input I, a counterfactual explanation "$I' > \neg R$" is given to the explainee, then we expect the counterfactual to be true in T_S—in formulas, $T_S \cup \{I\} \models (I' > \neg R)$—for some semantics of counterfactuals, which we do not want to delve into now.

Now, we consider the cases in which E is irrelevant:

Definition 1.

1. *(Irrelevance for the general context) E is irrelevant for Result R if there exists another input I', for which S yields a different result R', such that E would explain also the result R'*
2. *(Irrelevance for the specific explainee) E is irrelevant for explainee e if for all $i \in \{1, \ldots, n\}$ it holds $T_e \models E_i$, that is, no conjunct of E was unknown to the explainee*

We discuss the above definitions.

Point 1 considers irrelevant those explanations that are too general—that is, not cogent for the result to be explained. Consider for example a classification system, that outputs $R =$*Dolphin* when given as input the photo of some animal in the sea. The explanation $E =$*"Because it swims."* is irrelevant in this context, and could raise the request of a *contrastive explanation* [5]: *"Yes, but also sharks swim. Why did you say that this is a picture of a dolphin and not the one of a shark?"*. Observe that, in fact, E is truthful also for $R' =$*Shark* (presumably, for a different input picture I'). A relevant explanation, instead, would be *"Because the tail fin is horizontal."*

Point 2 takes into account the fact that explanations may be more than just one phrase, for instance when a chain of reasoning is shown, that leads from the input to the result. Point 2 requires that at least one conjunct E_i forming the explanation must be unknown to the user. Observe that we do not exclude parts of the explanation that are already known, since if they form an entire argument, excluding them would make the argument scattered. For instance, if the explanation for denying the loan was $E =$*"Because you are not resident in this country, and the risk-assessment threshold for non-residents is higher than the normal one"*, then $E = E_1 \wedge E_2$, where

$$
\begin{aligned}
E_1 &= \text{"you are not resident in this country"} \\
E_2 &= \text{"the risk-assessment threshold for non-residents is higher than the normal}
\end{aligned}
$$

one”

Now if e is not aware of E_2 (that can be checked as $T_e \not\models E_2$), then E as a whole may be considered relevant thanks of the presence of E_2. The presence of E_1 instead, although a little redundant, may be considered part of the entire argument, so E can be considered relevant even if E_1 is present. Note that in this brief discussion paper, we not tackle the question of *redundancy*, which we consider different from (ir)relevance.

2.1. More Examples

Let us think about two popular examples in explanation research: the arthropods classification by Miller [5] and the loan acceptance in the field of counterfactual explanations [6, 7].

2.1.1. Arthropod classification

Suppose that, given the classification of an image J, a user asks the question *"Why is image $\mathfrak{J}$ labelled as a spider instead of a beetle?"*. In this case, it is irrelevant for the classification context an explanation like *"Because it represents an arthropod"*; this explanation chunk, although true, is obviously true for all images and thus irrelevant for understanding the classification reasons.

Imagine now that the explainee is a biologist, asking to the contrastive explanation agent *"But an octopus can have eight legs too. Why did you not classify image $\mathfrak{J}$ as an octopus?"*. An explanation like $E = E_1 \wedge E_2$ =*"Because my function is only to classify arthropods, and an octopus is not an arthropod"* is relevant to a biologist only in its first part E_1. Instead, the information E_2 about the octopus category is well known by any biologist.

2.1.2. Loan granting

In the second example scenario, we focus on explanations of reasons for loan rejection. Any bank customer asking for a loan is not interested in rejection explanations like *“The risk associated to your loan request is too high"* (or, in a counterfactual fashion, *“If the risk associated to your loan request was lower, then the loan would have been accepted”*). This rejection condition is true for all rejected loan requests (the input-output pairs I'–R' of Point 1), and then irrelevant for the context. Customers would be much more interested in knowing their own specific reasons for risk evaluation: age, health conditions, income level, and so on.

Analogously, explaining to the customer *“You did not get the loan because you are over 40 years old"* is irrelevant, because tells something he/she already knows (his/her age). A relevant explanation might have been $E = E_1 \wedge E_2 \wedge E_3$, where

$$
\begin{aligned}
E_1 &= \text{“You are over 40 years old”} \\
E_2 &= \text{“The risk evaluated for customers over 40 years old is high”} \\
E_3 &= \text{“Loans are denied to high-risk factor applicants”}
\end{aligned}
$$

presuming that at least E_2 was unknown to the explainee.

2.1.3. Recommender Systems

The ability of explaining why a recommender system suggested a user a particular item (or set of items) is now recognized as an important feature [8]. In particular, counterfactual explanations of recommended items [9] may suggest the user alternative items that could be recommended, provided that the user's preferences change accordingly to the antecedent of the counterfactual. In symbols, a counterfactual explanation $p > q$ can be communicated to the user as *"I suggested you item s, but if your preferences were changed to p, I would suggest you item q instead of s"*.

In this application area, irrelevant explanations may hamper the user's trust in the recommendation system, obtaining an opposite effect explanations were devised for. Imagine a smartphone recommender system, a user entering preferences I, and being recommended a smartphone R. An example of counterfactual explanation being irrelevant for the context (Point 1 above) would be *"If you had no restrictions on budget, I would have suggested you an Apple iPhone 14 Pro 256GB."* While being true, such an explanation would fit any other preference setting I' and subsequent recommendation R', being the iPhone 14 Pro one of the possible obvious choices in case of unlimited budget. We note that researchers are implicitly aware of this kind of irrelevance, and usually, to avoid such explanations, explaining modules try to perturb as little as possible the initial input I (*e.g.*, raising the budget limits by a small amount only) in order to get a relevant counterfactual explanation, like for instance, *"If you raised your budget by 10$, I would have recommended you this other smartphone."*

3. Pruning explanation of irrelevant chunks: the RDF case study

The formalization above is not just a theoretical speculation on irrelevance. We applied the above criteria in a tool proposed by Colucci *et al.* [10, 11] to provide a human-readable explanation of commonalities shared by groups of RDF [12] resources, somehow aggregated (*e.g.*, by a clustering algorithm[2]).

The verbalization is based on the logic-based computation of the Least Common Subsumer(LCS) in RDF [14].

We apply the LCS-based verbalization tool to clustering results in two application scenarios: public procurement and drug comparison.

The first scenario is modelled in TheyBuyForYou (TBFY) dataset, a knowledge graph [15] that includes an ontology for procurement data, based on the Open Contracting Data Standard (OCDS)[16].

In particular, all contracting processes released on January, 30^{th} 2019 have been clustered with the K-means [17] algorithm[3] and the smallest cluster has been explained in terms of commonalities (on the basis of the LCS $R = L_1$ of the set of items I it contains).

The resulting explanation is given in Figure 1.

The reader may notice that the last explanation line—call it E_1—is objectively irrelevant in this context (so, for any user): any contracting process in the original dataset has been

[2]Note that the tool does not explain a whole partition into clusters of a set of resources, as in [13]; it only describes the commonalities of two or more resources already clusterized.

[3]The implementation at https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html has been used

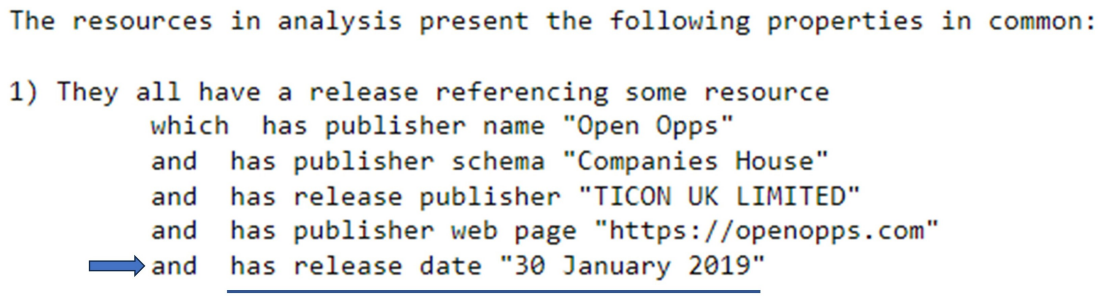

Figure 1: Explanation (obtained by the verbalization tool by Colucci *et al.* [10, 11] of the commonalities in the smallest cluster returned by clustering with k-Means all contracting processes released on January 30, 2019. The blue arrow indicates an irrelevant explanation chunk.

released on January 30, 2019, causing this explanation chunk to be obvious in the addressed clustering scenario. In terms of the previous formalization, we can automatically exclude E_1 in the following way: by computing the LCS of a wider set of resources I'—that is, adding to the cluster another random resource—we obtain an LCS $R' = L_2$ having, among others, the same release date already found in L_1. Since the explanation E_1 =*"Released on January 30, 2019"* is entailed by L_2, we can conclude that E_1 is irrelevant, and exclude it from the relevant explanations for L_1.

In the second application scenario, the search for similarities between drugs modelled in RDF is addressed. In particular, the *National Drug File - Reference Terminology* hosted by Bioportal[4] is used as a dataset.

Figure 2 shows an explanation for the similarity of two antibiotics: "cefepime" (*http://purl.bioontology.org/ontology/NDFRT/N0000022054*) and "ceftazidime" (*http://purl.bioontology.org/ontology/NDFRT/N0000145931*) produced by the verbalization tool by Colucci *et al.* [10, 11]).

If the explainee is a physician, some explanation chunks (blue arrows and lines in figure) are intuitively irrelevant: it is common knowledge (at least) for physicians that *(i)* any antibiotic may treat bacterial infections (and thus, infections) and that *(ii)* fever is a body temperature change.

The formalization we propose aims at pruning explanations of chunks which are irrelevant to the explainees if their knowledge is logically represented. In fact, by taking as T_e (among others) the RDFS statements expressing *(i)–(ii)* as a medical ontology in Bioportal, it is possible to automatically check that T_e entails[5] both *(i)* and *(ii)*, concluding that they are irrelevant for a physician, and exclude them from a concise explanation.

[4]https://bioportal.bioontology.org/ontologies/NDFRT

[5]In this case, entailment reduces to simple containment, but more elaborated examples involving blank nodes need entailment.

4. Conclusion

In this paper, we contribute to the discussion about explanation of AI-based tools, by formalizing a logical framework for identifying irrelevant chunks in a complex explanation.

Our proposal stems from the assumption that *irrelevance* is more amenable to a logical formalization than relevance, which is intrinsically subjective. In fact, we provide two definitions that may be implemented for pruning irrelevant portions of explanation: *i) irrelevance for the general context*: refers to knowledge true also for situations different from the one being explained; *ii) irrelevance for the specific explainee*: refers to knowledge already known to the explainee.

We demonstrate the practical applicability of these definitions, by implementing them in a tool that provides human-readable explanations of commonalities shared by group of RDF resources. Thanks to our formal definition, the use case shows how to prune complex similarity explanations by deleting irrelevant chunks.

Acknowledgements

We acknowledge partial support of the project "Contratto di Programma "Digital Enterprise", POR Puglia Grandi Imprese, Code NIL6S28.

References

[1] S. Kambhampati, Changing the nature of ai research, Commun. ACM 65 (2022) 8–9. URL: https://doi.org/10.1145/3546954. doi:10.1145/3546954.

[2] R. Tomsett, D. Braines, D. Harborne, A. D. Preece, S. Chakraborty, Interpretable to whom? A role-based model for analyzing interpretable machine learning systems, in: ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), 2018. URL: http://arxiv.org/abs/1806.07552. arXiv:1806.07552.

[3] J. Kim, A. Rohrbach, T. Darrell, J. Canny, Z. Akata, Textual explanations for self-driving vehicles, Proceedings of the European Conference on Computer Vision (ECCV) (2018).

[4] S. Colucci, F. M. Donini, E. D. Sciascio, On the relevance of explanation for RDF resources similarity, in: E. Babkin, M. Molhanec, P. Malyzhenkov, V. Merunka (Eds.), Proceedings of the 3rd Workshop on Model-driven Organizational and Business Agility 2023, co-located with the 35th International Conference on Advanced Information Systems Engineering (CAISE 2023), Zaragoza, Spain, June 12–16 2023, Lecture Notes in Business Information Processing, Springer, 2023. To appear.

[5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1 – 38. URL: http://www.sciencedirect.com/science/article/pii/S0004370218305988. doi:https://doi.org/10.1016/j.artint.2018.07.007.

[6] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology 31 (2018).

[7] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery Special Issue on Explainable and Interpretable Machine Learning and Data Mining (2022). doi:https://doi.org/10.1007/s10618-022-00831-6.

[8] Y. Zhang, X. Chen, Explainable recommendation: A survey and new perspectives, Found. Trends Inf. Retr. 14 (2020) 1–101. URL: https://doi.org/10.1561/1500000066. doi:10.1561/1500000066.

[9] J. Tan, S. Xu, Y. Ge, Y. Li, X. Chen, Y. Zhang, Counterfactual explainable recommendation, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 1784–1793. URL: https://doi.org/10.1145/3459637.3482420. doi:10.1145/3459637.3482420.

[10] S. Colucci, F. M. Donini, N. Iurilli, E. D. Sciascio, A business intelligence tool for explaining similarity, in: E. Babkin, J. Barjis, P. Malyzhenkov, V. Merunka (Eds.), Model-Driven Organizational and Business Agility - Second International Workshop, MOBA 2022, Leuven, Belgium, June 6-7, 2022, Revised Selected Papers, volume 457 of *Lecture Notes in Business Information Processing*, Springer, 2022, pp. 50–64. URL: https://doi.org/10.1007/978-3-031-17728-6_5. doi:10.1007/978-3-031-17728-6_5.

[11] S. Colucci, F. M. Donini, E. D. Sciascio, A human-readable explanation for the similarity of RDF resources, in: C. Musto, R. Guidotti, A. Monreale, G. Semeraro (Eds.), Proceedings of the 3rd Italian Workshop on Explainable Artificial Intelligence co-located with 21th International Conference of the Italian Association for Artificial Intelligence(AIxIA 2022), Udine, Italy, November 28 - December 3, 2022, volume 3277 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 88–103. URL: https://ceur-ws.org/Vol-3277/paper7.pdf.

[12] P. Patel-Schneider, D. Arndt, T. Haudebourg, RDF 1.2 semantics, https://www.w3.org/TR/rdf12-semantics/, 2023.

[13] M. Moshkovitz, S. Dasgupta, C. Rashtchian, N. Frost, Explainable k-means and k-medians clustering, in: H. D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 7055–7065. URL: https://proceedings.mlr.press/v119/moshkovitz20a.html.

[14] S. Colucci, F. Donini, S. Giannini, E. Di Sciascio, Defining and computing least common subsumers in RDF, Web Semantics: Science, Services and Agents on the World Wide Web 39 (2016) 62 – 80.

[15] A. Soylu, O. Corcho, B. Elvesater, C. Badenes-Olmedo, T. Blount, F. Yedro Martinez, M. Kovacic, M. Posinkovic, I. Makgill, C. Taggart, E. Simperl, T. C. Lech, D. Roman, TheyBuyForYou platform and knowledge graph: Expanding horizons in public procurement with open linked data, Semantic Web 13 (2022) 265–291.

[16] A. Soylu, B. Elvesæter, P. Turk, D. Roman, O. Corcho, E. Simperl, G. Konstantinidis, T. C. Lech, Towards an ontology for public procurement based on the open contracting data standard, Springer-Verlag, Berlin, Heidelberg, 2019, p. 230–237.

[17] X. Jin, J. Han, K-Means Clustering, Springer US, Boston, MA, 2010, pp. 563–564. URL: https://doi.org/10.1007/978-0-387-30164-8_425. doi:10.1007/978-0-387-30164-8_425.

1)They are all Organic Chemical that is Chemical Viewed Structurally ;

2)They are all Antibiotic that is Pharmacologic Substance;

3)They are all C preparations that is Classification;

4)They are all C preparations that is Drug Products by Generic Ingredient Combinations;

5)They all has_ingredient something that is Organic Chemical that is Chemical Viewed Structurally;

6)They all has_ingredient something that is Antibiotic that is Pharmacologic Substance;

7)They all has_physiologic_effect Decreased Cell Wall Synthesis & Repair that is Organ or Tissue Function;

8)They all has_physiologic_effect Decreased Cell Wall Synthesis & Repair that is Cell Wall Alteration;

9)They all may_treat Serratia Infections that is Enterobacteriaceae Infections;

10)They all may_treat something that is Enterobacteriaceae Infections;

11)They all may_treat Urinary Tract Infections that is Infection;

12)They all may_treat something that is Infection;

13)They all may_treat Acinetobacter Infections that is Moraxellaceae Infections;

14)They all may_treat Escherichia coli Infections that is Enterobacteriaceae Infections;

15)They all may_treat something that is Enterobacteriaceae Infections;

16)They all may_treat Neutropenia that is Agranulocytosi;

17)They all may_treat Pneumonia, Bacterial that is Bacterial Infections;

18)They all may_treat something that is Bacterial Infections;

19)They all may_treat Haemophilus Infections that is Pasteurellaceae Infections;

20)They all may_treat Streptococcal Infections that is Gram-Positive Bacterial Infections ;

21)They all may_treat something that is Gram-Positive Bacterial Infections ;

22)They all may_treat Proteus Infections that is Enterobacteriaceae Infection;

23)They all may_treat Fever that is Finding ;

24)They all may_treat Fever that is Body Temperature Changes;

25)They all may_treat Fever induced_by PLAGUE VACCINE INJ ;

26)They all may_treat Sepsis that is Infection;

27)They all may_treat Pseudomonas Infections that is Gram-Negative Bacterial Infections;

28)They all may_treat Klebsiella Infections that is Enterobacteriaceae Infections ;

29)They all may_treat Bone Diseases, Infectious that is Bone Diseases;

30)They all may_treat Skin Diseases, Bacterial that is Bacterial Infections;

31)They all are contraindicated with Drug Hypersensitivity that is Hypersensitivity ;

32)They all has_mechanism_of_action Enzyme Inhibitors that is Molecular Function;

33)They all share has_mechanism_of_action Enzyme Inhibitors that is Enzyme Interactions ;

Figure 2: Explanation of the commonalities between "cefepime" (*http://purl.bioontology.org/ontology/NDFRT/N0000022054*) and "ceftazidime" (*http://purl.bioontology.org/ontology/NDFRT/N0000145931*) computed through the verbalization tool by Colucci *et al.* [10, 11]. The blue arrows and lines indicate irrelevant explanation chunks.