Can You Tell the Difference? Contrastive Explanations for ABox Entailments (Extended Abstract)

Patrick Koopmann[1], Yasir Mahmood[2] and Axel-Cyrille Ngonga Ngomo[2]

[1]Knowledge in Artificial Intelligence, Vrije Universiteit Amsterdam, The Netherlands
[2]Data Science Group, Heinz Nixdorf Institute, Paderborn University, Germany

Keywords
Description Logic, Explanations, Computational Complexity, Abduction, Justifications

Due to the expressive power of description logics (DLs), and the complexity of real ontologies, inferences performed by a DL reasoner are not always easy to understand by end-users, which motivates a variety of research projects on explaining entailments in DL knowledge bases (KBs). Such research is usually either concerned with *why*-questions ("Why is X entailed by the KB?") or with *why not*-questions ("Why is X *not* entailed by the KB?"). Approaches for answering the why-question include *justifications* (i.e. minimal subsets of the KB that are sufficient for the entailment [1, 2, 3, 4]), proofs [5], Craig interpolants [6] and universal models [7]. To answer a *why not* question, we can use *abductive reasoning* to determine what is missing in a KB $\mathcal{K}$ to derive an observation α [8, 9, 10]. Research in this area for DLs encompasses *ABox abduction* [11, 12], *TBox abduction* [13], *KB abduction* [14] and *concept abduction* [15], depending on the type of entailment to be explained.

We are interested on explaining outcomes of instance queries, i.e. explaining entailments and non-entailments of assertions of the form $C(a)$. In this context, addressing the *why* and *why not* questions jointly can provide greater clarity than considering them in isolation. To illustrate this, consider a simplified KB $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ for a hiring process:

$$\begin{aligned}\mathcal{T} = \{\ & \mathit{Qualified} \sqcap \exists \mathit{publishedAt}.\mathit{Journal} \sqsubseteq \mathit{Interviewed}, \qquad \mathit{PostDoc} \sqcap \exists \mathit{leads}.\mathit{Group} \sqsubseteq \bot \\ & \exists \mathit{leads}.\mathit{Group} \sqcup \exists \mathit{hasFunding}.\top \sqsubseteq \mathit{Qualified} \quad \} \\ \mathcal{A} = \{\ & (1)\ \mathit{publishedAt}(\mathit{alice}, \mathit{jair}), \quad (2)\ \mathit{publishedAt}(\mathit{bob}, \mathit{kr}), \quad (3)\ \mathit{Journal}(\mathit{jair}), \\ & (4)\ \mathit{leads}(\mathit{alice}, \mathit{cs}), \quad (5)\ \mathit{Group}(\mathit{cs}), \quad (6)\ \mathit{hasFunding}(\mathit{alice}, \mathit{corp}), \quad (7)\ \mathit{PostDoc}(\mathit{bob})\}\end{aligned}$$

We can explain why Alice was interviewed with the ABox justification $\{(1), (3), (4), (5)\}$. Using ABox abduction, we can explain why Bob is not interviewed using the hypothesis: $\{\ \mathit{Journal}(\mathit{kr}),\ \mathit{hasFunding}(\mathit{bob}, g)\ \}$ for fresh g. But to explain why Alice got invited and Bob did not, a more on point explanation would minimize the differences: we may want to focus on the fact that Alice receives funding while Bob does not, rather than that she also leads a group, which Bob cannot. Specifically, we integrate both explanations into one *contrastive explanation.*

Lipton (1990) introduced the notion of contrastive explanations with the goal to express an inquirer's *preference*, or reflect their demand regarding the *context* in which an explanation is requested (e.g., explain why bob was not interviewed, *in the context* of alice, who was). Contrastive explanations have also been considered for answer set programming [17] as well as in explaining classification of machine learning models [18, 19, 20, 21, 22]. Another related term is *counter-factual explanations* in sub-symbolic machine learning [23, 24]. What these approaches have in common is that they look at similarities and differences at the same time, and that they use syntactic *patterns* to highlight the differences. We argue that in the context of ABox reasoning, patterns should reflect the structure of the ABox using *ABox patterns* (which are essentially conjunctive queries), that are instantiated differently for the fact and the foil.

DL 2025: 38th International Workshop on Description Logics, September 3–6, 2025, Opole, Poland

p.k.koopmann@vu.nl (P. Koopmann); yasir.mahmood@uni-paderborn.de (Y. Mahmood); axel.ngonga@upb.de (A. N. Ngomo)
0000-0001-5999-2583 (P. Koopmann); 0000-0002-5651-5391 (Y. Mahmood); 0000-0001-7112-3516 (A. N. Ngomo)

Contributions. We propose a framework for contrastive ABox explanations (CEs) in DLs. We distinguish between a syntactic and a semantic version, consider different optimality criteria (minimizing differences and conflicts, maximizing similarities), and analyze them for different DLs ranging from $\mathcal{EL}$ to $\mathcal{ALCI}$. We develop a first prototype for computing difference-minimal explanations and evaluate it on ORE 2015 ontologies.

Our Definition. A *contrastive ABox explanation problem* (CP) is a tuple $P = \langle \mathcal{K}, C, a, b\rangle$ consisting of a KB $\mathcal{K}$, a concept C and two individual names a, b, s.t. $\mathcal{K} \models C(a)$ and $\mathcal{K} \not\models C(b)$. Intuitively, P reads as *"Why is a an instance of C and b is not?"*. We call a (or $C(a)$) the *fact* and b ($C(b)$) the *foil* of the CP. Note that in a CP, $\mathcal{K}$ is always consistent as $\mathcal{K} \not\models C(b)$.

Following Lipton (1990), we contrast a and b by highlighting the *differences* between the reasons that support $C(a)$ and the *missing* elements that would support $C(b)$. Since different individuals may be related to a than to b, we abstract away from concrete individual names and use instead *ABox patterns*. Here, an ABox pattern is a set $q(\vec{x})$ of ABox assertions that uses variables from $\vec{x}$ instead of individual names. Given a vector $\vec{c}$ of individual names with the same length as $\vec{x}$, $q(\vec{c})$ denotes the ABox assertions obtained after replacing variables by individuals according to $x_i \mapsto c_i$. The goal is to characterize the *difference* between individuals a and b using an ABox pattern $q_{\mathit{diff}}(\vec{x})$, paired with vectors $\vec{c}$ and $\vec{d}$ such that $q_{\mathit{diff}}(\vec{c})$ is entailed by the KB, $q_{\mathit{diff}}(\vec{d})$ is not, and adding $q_{\mathit{diff}}(\vec{d})$ to the KB would entail $C(b)$. In our running example, $q(x, y, z) = \{\mathit{Journal}(y), \mathit{hasFunding}(x, z)\}$ is such an ABox pattern, where for the fact *alice* we have $\vec{c} = \langle \mathit{alice}, \mathit{jair}, \mathit{corp}\rangle$, and for the foil *bob* we use $\langle \mathit{bob}, \mathit{kr}, g\rangle$, for a fresh individual *g*.

To really *explain* the (missing) entailment, we also have to include other assertions relevant to achieve the missing entailment. In our example, understanding the explanation also requires knowing the *commonality* $q_{\mathit{com}} = \{\mathit{publishedAt}(x, y)\}$ between the fact and foil. Precisely, our contrastive explanations use ABox patterns $q(\vec{x}) = q_{\mathit{com}}(\vec{x}) \cup q_{\mathit{diff}}(\vec{x})$, with $q_{\mathit{com}}(\vec{x})$ referring to what holds for both instantiations, and $q_{\mathit{diff}}(\vec{x})$ what only holds for the fact. Finally, in certain cases it might not be possible to obtain an explanation without triggering an inconsistency. Thus, we also permit difference-parts to be conflicting with the KB. To see this, let us replace the axiom "$\exists \mathit{leads}.\mathit{Group} \sqcup \exists \mathit{hasFunding}.\top \sqsubseteq \mathit{Qualified}$" in our example by "$\exists \mathit{leads}.\mathit{Group} \sqsubseteq \mathit{Qualified}$". We might still want to provide an explanation, for instance: "If KR was a journal and Bob lead the CS group, he would have been interviewed, but being a postdoc, he cannot lead a group." To consider such explanations, we add a final component called the *conflict* set in our definition. Removing such conflicts from KB results in an *alternative* (*counterfactual*) scenario consistent with the proposed explanation. These observations result in the following definition:

Definition 1. *Let $P = \langle \mathcal{K}, C, a, b\rangle$ be a CP where $\mathcal{K} = \langle \mathcal{T}, \mathcal{A}\rangle$. A solution to P (the* contrastive ABox explanation / CE*) is a tuple $\langle q_{\mathit{com}}(\vec{x}), q_{\mathit{diff}}(\vec{x}), \vec{c}, \vec{d}, \mathcal{C}\rangle$ of ABox patterns $q_{\mathit{com}}(\vec{x})$, $q_{\mathit{diff}}(\vec{x})$, vectors $\vec{c}$ and $\vec{d}$ of individual names, and a set $\mathcal{C}$ of ABox assertions, which for $q(\vec{x}) = q_{\mathit{com}}(\vec{x}) \cup q_{\mathit{diff}}(\vec{x})$ satisfies the following conditions:*

1. *$\mathcal{T}, q(\vec{c}) \models C(a)$ and $\mathcal{T}, q(\vec{d}) \models C(b)$,*
2. *$\mathcal{K} \models q(\vec{c})$,*
3. *$\mathcal{K} \models q_{\mathit{com}}(\vec{d})$,*
4. *$q(\vec{c})$ is a $\subseteq$-minimal set satisfying the conditions 1 and 2,*
5. *$\mathcal{C} \subseteq \mathcal{A}$ is $\subseteq$-minimal such that $\mathcal{T}, (\mathcal{A} \setminus \mathcal{C}) \cup q(\vec{d}) \not\models \bot$.*

We call $\vec{c}$ and $\vec{d}$ the *fact evidence* and the *foil evidence*. The patterns $q_{\mathit{com}}(\vec{x})$ and $q_{\mathit{diff}}(\vec{x})$ will be called the *commonality* and the *difference* between a and b. Intuitively, $q(\vec{x}) = q_{\mathit{com}}(\vec{x}) \cup q_{\mathit{diff}}(\vec{x})$ describes a pattern that is responsible for a being an instance of C, with $q_{\mathit{com}}(\vec{x})$ describing what a and b have in common, and $q_{\mathit{diff}}(\vec{x})$ what b is lacking. By instantiating $\vec{x}$ with $\vec{c}$ we obtain a set of ABox axioms that entails $C(a)$, and by instantiating it with $\vec{d}$, we obtain a set of axioms that would entail $C(b)$, where $q_{\mathit{com}}(\vec{d})$ is already provided by the present ABox, and $q_{\mathit{diff}}(\vec{d})$ is missing. Since, $q_{\mathit{diff}}(\vec{d})$ can be inconsistent with the KB, the set $\mathcal{C}$ presents the conflicts and $(\mathcal{A} \setminus \mathcal{C}) \cup q(\vec{d})$ depicts an alternative (consistent) scenario in which $C(b)$ is true.

	fresh individuals	$\mathcal{EL}_\bot$		$\mathcal{ALC}/\mathcal{ALCI}$	
		$\subseteq$	$\leq$	$\subseteq$	$\leq$
difference-minimal	–	P	coNP	ExpTime	ExpTime
conflict-minimal	yes	ExpTime	ExpTime	coNExpTime	coNExpTime
	no	coNP	coNP	ExpTime	ExpTime
commonality-maximal	–	–	coNP	ExpTime	ExpTime

Table 1
Complexity of verification for syntactic CEs.

Example 1. *For the example in the introduction, the CP is:* $\langle \mathcal{K}, \mathit{Interviewed}, \mathit{alice}, \mathit{bob}\rangle$. *A CE for this CP is* $\langle q_{\mathit{com}}(x,y,z),\ q_{\mathit{diff}}(x,y,z),\ \vec{c},\ \vec{d},\ \mathcal{C}\rangle$, *where*

$$q_{\mathit{com}}(x,y,z) = \{\mathit{publishedAt}(x,y)\}, \qquad q_{\mathit{diff}}(x,y,z) = \{\mathit{hasFunding}(x,z),\ \mathit{Journal}(y)\}$$

$$\vec{c} = \langle \mathit{alice}, \mathit{jair}, \mathit{corp}\rangle, \qquad \vec{d} = \langle \mathit{bob}, \mathit{kr}, g\rangle, \qquad \mathcal{C} = \emptyset$$

Variants of CEs. A natural restriction to Definition 1 is to limit it to *syntactic CEs.* For a CP $P = \langle \mathcal{K}, C, a, b\rangle$ with KB $\mathcal{K} = \langle \mathcal{T}, \mathcal{A}\rangle$, the CE $E = \langle q_{\mathit{com}}(\vec{x}), q_{\mathit{diff}}(\vec{x}), \vec{c}, \vec{d}, \mathcal{C}\rangle$ is *syntactic* if $q_{\mathit{com}}(\vec{c}), q_{\mathit{diff}}(\vec{c}), q_{\mathit{com}}(\vec{d}) \subseteq \mathcal{A}$. Syntactic explanations can only refer to what is explicit in the ABox while semantic explanations can also refer to implicitly entailed information. Semantic explanations can be useful to highlight the commonalities and differences more specifically.

Optimality Criteria. Condition 4 in Definition 1 requires the pattern $q(\vec{c})$ to be subset-minimal and helps to avoid redundant elements in the explanation. In addition, it seems natural to require both the difference and the conflict to be as small as possible, while wanting to maximize the commonality-part. Finally, optimality may be defined locally (e.g. subset-minimal) or globally (cardinality-maximal). This leads to the following optimality criteria.

Given a CP P and a contrastive explanation $E = \langle q_{\mathit{com}}(\vec{x}), q_{\mathit{diff}}(\vec{x}), \vec{c}, \vec{d}, \mathcal{C}\rangle$ for P. Then,

- E is *conflict-minimal* if no explanation E' has a conflict set $\mathcal{C}' \subset \mathcal{C}$.
- E is *commonality-maximal* if no explanation E' has commonality $q'_{\mathit{com}}(\vec{x}')$ and foil evidence $\vec{d}'$, such that $q_{\mathit{com}}(\vec{d}) \subset q'_{\mathit{com}}(\vec{d}')$.
- E is *difference-minimal* if no explanation E' has difference $q'_{\mathit{diff}}(\vec{x}')$ and foil evidence $\vec{d}'$, such that $q'_{\mathit{diff}}(\vec{d}') \subset q_{\mathit{diff}}(\vec{d})$.

We further define each of the aforementioned optimality w.r.t. the cardinality of given sets.

Complexity Results. We note that the *existence problem* is often uninteresting: syntactic CE candidates can be easily constructed from a justification for the fact, and thus always exist. Furthermore, the existence of a solution also implies the existence of a subset- or cardinality-minimal one for each component one prefers. Therefore, from the complexity theoretic perspective, a more interesting problem is the verification of a CE when given as input. Given a CP P formulated in DL $\mathcal{L}$ and an explanation E for P, determine whether E is a valid D-CE for P with optimal C. We considered DLs $\mathcal{L} \in \{\mathcal{EL}_\bot, \mathcal{ALC}, \mathcal{ALCI}\}$, definitions $D \in \{\text{syntactic}, \text{semantic}\}$ and optimal criteria $C \in \{\text{conflict}, \text{commonality}, \text{difference}\}$. An interesting observation is that conflict-minimality may require exponentially many fresh individuals in the foil vector, thus leading to a high complexity. We thus also consider the case without fresh individuals here.

Prototypical Implementation. We show that difference-minimal CEs can be computed in polynomial time with an oracle deciding logical entailment, while all the other criteria are intractable. Based on this observation, we developed a first prototype to compute difference-minimal syntactic CEs. We

Corpus	#CPs		Commonality		Difference		Conflict		Fresh Individuals		Duration (sec.)	
	average	range	average	range	average	range	average	range	average	range	average	range
$\mathcal{EL}_\bot$	35.1	4 – 50	0.45	0 – 4	1.42	1 – 6	0.0	0 – 0	0.34	0 – 3	2.84	0.08 – 386.9
$\mathcal{ALCI}$	34.7	1 – 50	0.34	0 – 7	1.29	1 – 5	0.36	0 – 9	0.25	0 – 5	8.81	0.06 – 493.6

Table 2
Experimental results for $\mathcal{EL}_\bot$ and $\mathcal{ALCI}$ corpus. "#CP" states the number of CPs answered (out of 50) within the timeout of 10 minutes. The other columns give statistics about the individual CEs computed.

evaluated it on ontologies from the OWL Reasoner Competition ORE 2015 [25, 26], restricted to the fragments we support, and using randomly generated explanation problems with concepts of size 5. Results are shown in Table 2—though not optimized much, we could often compute CEs in practice, and all the elements allowed by our definition occurred in computed CEs. Our results indicate that the CEs computed tended to be simple, despite the concept to be explained of size five.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] F. Baader, B. Hollunder, Embedding defaults into terminological knowledge representation formalisms, Journal of Automated Reasoning 14 (1995) 149–180.

[2] S. Schlobach, R. Cornet, et al., Non-standard reasoning services for the debugging of description logic terminologies, in: Ijcai, volume 3, 2003, pp. 355–362.

[3] A. Kalyanpur, Debugging and repair of OWL ontologies, University of Maryland, College Park, 2006.

[4] M. Horridge, Justification based explanation in ontologies, The University of Manchester (United Kingdom), 2011.

[5] C. Alrabbaa, F. Baader, S. Borgwardt, P. Koopmann, A. Kovtunova, Finding good proofs for description logic entailments using recursive quality measures, in: A. Platzer, G. Sutcliffe (Eds.), Automated Deduction - CADE 28 - 28th International Conference on Automated Deduction, Virtual Event, July 12-15, 2021, Proceedings, volume 12699 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 291–308. URL: https://doi.org/10.1007/978-3-030-79876-5_17. doi:10.1007/978-3-030-79876-5_17.

[6] S. Schlobach, Explaining subsumption by optimal interpolation, in: J. J. Alferes, J. A. Leite (Eds.), Logics in Artificial Intelligence, 9th European Conference, JELIA 2004, Lisbon, Portugal, September 27-30, 2004, Proceedings, volume 3229 of *Lecture Notes in Computer Science*, Springer, 2004, pp. 413–425. URL: https://doi.org/10.1007/978-3-540-30227-8_35. doi:10.1007/978-3-540-30227-8_35.

[7] C. Alrabbaa, S. Borgwardt, P. Koopmann, A. Kovtunova, Explaining ontology-mediated query answers using proofs over universal models, in: International Joint Conference on Rules and Reasoning, Springer, 2022, pp. 167–182.

[8] C. Peirce, Deduction, induction and hypothesis: Popular science monthly, v. 13 (1878).

[9] C. Elsenbroich, O. Kutz, U. Sattler, A case for abductive reasoning over ontologies, in: Proceedings of the OWLED 2006 Workshop on OWL: Experiences and Directions, Athens, Georgia, USA, November 10-11, 2006, volume 216, CEUR, 2006, pp. 1–12.

[10] F. Wei-Kleiner, Z. Dragisic, P. Lambrix, Abduction framework for repairing incomplete $\mathcal{EL}$ ontologies: Complexity results and algorithms, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 28, 2014.

[11] S. Klarman, U. Endriss, S. Schlobach, Abox abduction in the description logic, Journal of Automated Reasoning 46 (2011) 43–80.

[12] W. Del-Pinto, R. A. Schmidt, ABox abduction via forgetting in $\mathcal{ALC}$, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, 2019, pp. 2768–2775.

[13] J. Du, H. Wan, H. Ma, Practical TBox abduction based on justification patterns, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, 2017, pp. 1100–1106. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14402.

[14] P. Koopmann, W. Del-Pinto, S. Tourret, R. A. Schmidt, Signature-based abduction for expressive description logics, in: D. Calvanese, E. Erdem, M. Thielscher (Eds.), Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020, 2020, pp. 592–602. URL: https://doi.org/10.24963/kr.2020/59. doi:10.24963/KR.2020/59.

[15] B. Glimm, Y. Kazakov, M. Welt, Concept abduction for description logics, in: Proceedings of the 35th International Workshop on Description Logics (DL 2022) co-located with Federated Logic Conference (FLoC 2022), Haifa, Israel, August 7th to 10th, 2022, 2022. URL: https://ceur-ws.org/Vol-3263/paper-11.pdf.

[16] P. Lipton, Contrastive explanation, Royal Institute of Philosophy Supplements 27 (1990) 247–266.

[17] T. Eiter, T. Geibinger, J. Oetsch, Contrastive explanations for answer-set programs, in: European Conference on Logics in Artificial Intelligence, Springer, 2023, pp. 73–89.

[18] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, P. Das, Explanations based on the missing: Towards contrastive explanations with pertinent negatives, Advances in neural information processing systems 31 (2018).

[19] A. Ignatiev, N. Narodytska, N. Asher, J. Marques-Silva, From contrastive to abductive explanations and back again, in: AIxIA 2020 - Advances in Artificial Intelligence: XIX Int. Conf. of the Italian Association for AI, volume 12414, Springer, 2020, pp. 335–355.

[20] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001.

[21] T. Miller, Contrastive explanation: a structural-model approach, The Knowledge Engineering Review 36 (2021) e14. doi:10.1017/S0269888921000102.

[22] J. Marques-Silva, A. Ignatiev, Delivering trustworthy AI through formal XAI, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI Press, 2022, pp. 12342–12350.

[23] S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, arXiv preprint arXiv:2010.10596 2 (2020) 1.

[24] S. Dandl, C. Molnar, M. Binder, B. Bischl, Multi-objective counterfactual explanations, in: International conference on parallel problem solving from nature, Springer, 2020, pp. 448–469.

[25] N. Matentzoglu, B. Parsia, ORE 2015 reasoner competition corpus, Dataset on Zenodo, 2015. URL: https://doi.org/10.5281/zenodo.18578. doi:10.5281/zenodo.18578.

[26] B. Parsia, N. Matentzoglu, R. S. Gonçalves, B. Glimm, A. Steigmiller, The OWL reasoner evaluation (ORE) 2015 competition report, J. Autom. Reason. 59 (2017) 455–482. URL: https://doi.org/10.1007/s10817-017-9406-8. doi:10.1007/S10817-017-9406-8.