# Negative Statements Considered Useful (Extended Abstract)

**Hiba Arnaout**[1] , **Simon Razniewski**[1] , **Gerhard Weikum**[1] , **Jeff Z. Pan**[2]

[1]Max Planck Institute for Informatics, Germany

[2]The University of Edinburgh, United Kingdom

{harnaout, srazniew, weikum}@mpi-inf.mpg.de, j.z.pan@ed.ac.uk

| Salient Negative Triple | Provenance | Score |
|---|---|---|
| ¬ (Abraham Lincoln, death cause, natural) | Unlike the previous 17 U.S. presidents. | 1.12 |
| ¬ (Jeff Bezos, occupation, politician) | Unlike the previous 17 of 21 Time Person of the Year winners. | 1.02 |
| ¬ (Angela Merkel, gender, male) | Unlike the previous 6 Chairmen of the CDU. | 0.89 |
| ¬ (Paul McCartney, citizenship, U.S.A.) | Unlike the previous 35 of 41 Grammy Award winners. | 1.20 |

**Abstract**

In this extended abstract, we summarize the main contributions of our recent work on identifying *salient negative* triples in open-world knowledge graphs (Arnaout et al. 2021).

## 1 Introduction

Structured knowledge is crucial in a range of applications like question answering (He et al. 2023; Chen et al. 2021; Fawei et al. 2019) and dialogue agents. The required knowledge is usually stored in Knowledge Graphs (KGs) (Pan et al. 2017), with notable projects like Wikidata (Vrandečić and Krötzsch 2014) and Yago (Suchanek, Kasneci, and Weikum 2007). At present, most major KGs only contain positive statements, e.g., *"Tom Cruise is an actor"*, whereas statements such as that *"Tom Cruise did not win an Oscar"* could only be inferred with the major assumption that the KG is complete (i.e., the *closed-world assumption* CWA (Reiter 1981; Ren, Pan, and Zhao 2010)). Yet as KGs are only pragmatic collections of positive statements, the CWA is not realistic to assume, and there remains uncertainty whether statements not contained in a KG are false, or their truth is merely unknown to the KG.

In this work, we make the case that *important* negative knowledge should be explicitly materialized. We motivate this selective materialization with the challenge of overseeing a large space of false statements, and with the importance of explicit negation in search and question answering (QA), that are mainly geared for positive questions. For instance, for answering negative questions like *"Actors without Oscars"*, QA systems lack a data basis. Similarly, they struggle with positive questions that have no answer, like *"Children of Emmanuel Macron"*, too often still returning a best-effort answer even if it is incorrect. Materialized negative information would allow a better treatment of both cases.

In this paper, we consider three classes of negative triples:

1. Grounded negative triples, e.g., ¬ (Tom Cruise, citizenship, U.K.), i.e., *Tom Cruise is not British*.
2. Universally absent negative triples, e.g., $\neg\exists o$ :(Tom Cruise, political party, o), i.e., *Tom Cruise is not a member of any political party*.
3. Conditional negative triples, e.g., ¬ (Tom Cruise, award, o) . (o, instance of, Oscar), i.e., *Tom Cruise has not won an award from the Oscar categories*.

To identify salient negative triples about an input entity, we propose the *peer-based negation inference* method. First, we propose several measures to compute highly related entities for almost any existing input entity, e.g., by measuring cosine similarity of pre-trained entity embeddings (Yamada et al. 2020). For instance, for the input entity Stephen Hawking, closest peers include other physicists such as Max Planck and Albert Einstein. These are later used to define parts of the KG where completeness is postulated, i.e. the LCWA (Local Closed-world Assumption). In this example, positive triples about peer entities such as (Max Planck, award, Nobel Prize in Physics) become candidate negative triples for Stephen Hawking (assuming they are not asserted for Hawking). To ensure the correctness of candidates, we exploit the PCA (Partial Completeness Assumption (Galárraga et al. 2013)). The PCA is one instantiation of the LCWA, which asserts that if a subject has *at least one* object for a given relation, then there are no other objects beyond those that are in the KG. For instance, if Stephen Hawking has at least one citizenship in a KG like Wikidata, we make the assumption that his list of citizenships is complete. The rationale behind this assumption is that salient relations, such as "has child" or "citizen of", especially for prominent entities, will either be covered completely, throughout the KG construction and maintenance process, or not at all. To sort potentially large sets of candidate triples, we finally propose four ranking metrics, combin-

ing frequency signals with popularity and probabilistic likelihoods, in a learn-to-rank model, to measure the salience of candidate negative statements. For instance, it is more salient that *Stephen Hawking never received a Nobel Prize in Physics* than than *He is not German*.

In an extension of this method, the *order-oriented peer-based negation inference*, we explore the usage of ordered lists of peers, which shows an improvement in salience. In contrast to the previous method, we consider the degree of peerness as an important element in determining the salience of a candidate negation. For instance, given the input entity Jeff Bezos, one possible ordered peer group is other winners of the Time Person of the Year Award. The members are ordered by the *date of their win*, allowing for provenance-extended negative statements such as *Jeff Bezos is not a politician, unlike the previous 17 out of 21 winners of the Time Person of the Year award.*

Moreover, we define the notion of *conditional negative statements* to express complex negation. For instance, the conditional negation that *Einstein did not study at any U.S. university* can be observed by negating, for each U.S. university in the KG, that he did not study there, resulting in potentially hundreds of grounded negatives that can only make the point when joined together. A more practical way of expressing this is to allow *conditional negatives* that can summarize and lift such cases. We present an algorithm to lift them from previously inferred simple ones.

Finally, we publish *Wikinegata*, a system to showcase the peer-based negation inference method, where users can explore negatives about 500K Wikidata entities, from 11 classes, while adjusting different parameters of the method, such as the negative triple's type and peering function.

The demo system can be accessed at:
https://d5demos.mpi-inf.mpg.de/negation/

The salient contributions of the paper are:

- We present judiciously designed methods for collecting and ranking interesting and explainable negative statements based on knowledge about highly related entities.
- We present a method to construct (complex) conditional negative statements from (simple) grounded ones.
- We show the usefulness of our model in use cases like entity summarization, decision support, and question answering.
- We release the first datasets of useful negative KG triples, and a platform to query them.

## 2 Relevance to KR

RDF (Pan 2009) KGs is a KR topic, and previous KR work argues in favor of explicit negation (Analyti et al. 2013), by proposing extended RDF, where an ERDF triple can be either positive or negative. Our work increases the useability of KGs by augmenting them with *valuable* negatives.

In this work, we use KR related techniques as building blocks to ensure the correctness of candidate negatives. In particular, we predict the completeness for certain KG relations based on the PCA rule (Galárraga et al. 2013), which significantly improves the correctness of our inferred candidates.

Existing methods (Galárraga et al. 2017; Ortona, Meduri, and Papotti 2018) employed these rules to identify correct negatives. However, our work is the first to combine this effort with heuristics to compile lists of, not only correct, but also *useful* negative triples.

In addition, we envision that the community on representation learning of KGs will find our large-scale datasets and methods to be a better way for negative sampling, e.g., in applications such as link prediction.

## References


Analyti, A.; Antoniou, G.; Damásio, C. V.; and Pachoulakis, I. 2013. A framework for modular ERDF ontologies. *Annals of Mathematics and Artificial Intelligence*.

Arnaout, H.; Razniewski, S.; Weikum, G.; and Pan, J. Z. 2021. Negative statements considered useful. *Journal of Web Semantics* 71:100661.

Chen, Z.; Chen, J.; Geng, Y.; Pan, J. Z.; Yuan, Z.; and Chen, H. 2021. Zero-Shot Visual Question Answering Using Knowledge Graph. In *Proc. of ISWC)*, 146–162.

Fawei, B.; Pan, J. Z.; Kollingbaum, M. J.; and Wyner, A. Z. 2019. A Semi-automated Ontology Construction for Legal Question Answering. *New Generation Computing* 453–478.

Galárraga, L.; Teflioudi, C.; Hose, K.; and Suchanek, F. 2013. Amie: Association rule mining under incomplete evidence in ontological knowledge bases. In *The Web Conf.*

Galárraga, L.; Razniewski, S.; Amarilli, A.; and Suchanek, F. M. 2017. Predicting completeness in knowledge bases. In *International Conference on Web Search and Data Mining*.

He, J.; U, S. C. L.; Gutiérrez-Basulto, V.; and Pan, J. Z. 2023. BUCA: A Binary Classification Approach to Unsupervised Commonsense Question Answering. In *ACL*.

Ortona, S.; Meduri, V. V.; and Papotti, P. 2018. RuDiK: rule discovery in knowledge bases. *Conference on Very Large Data Bases*.

Pan, J. Z.; Vetere, G.; Gomez-Perez, J. M.; and Wu, H., eds. 2017. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.

Pan, J. Z. 2009. Resource Description Framework. In *Handbook of Ontologies*.

Reiter, R. 1981. On closed world data bases. In *Readings in artificial intelligence*. Elsevier.

Ren, Y.; Pan, J. Z.; and Zhao, Y. 2010. Closed World Reasoning for OWL2 with NBox. *Journal of Tsinghua Science and Technology* (6).

Suchanek, F. M.; Kasneci, G.; and Weikum, G. 2007. Yago: A core of semantic knowledge. In *The Web Conference*.

Vrandečić, D., and Krötzsch, M. 2014. Wikidata: a free collaborative knowledge base. *Communications of the ACM*.

Yamada, I.; Asai, A.; Sakuma, J.; Shindo, H.; Takeda, H.; Takefuji, Y.; and Matsumoto, Y. 2020. Wikipedia2Vec: an optimized tool for learning embeddings of words and entities from wikipedia. *EMNLP*.