Relation Canonicalization in Open Knowledge Graphs: A Quantitative Analysis

Maria Lomaeva and Nitisha Jain[0000−0002−7429−7949]

Hasso Plattner Institute
University of Potsdam, Potsdam, Germany
lomaeva@uni-potsdam.de, nitisha.jain@hpi.de

Abstract. Open Information Extraction (OpenIE) allows the detection of meaningful triples of *(noun phrase, relation phrase, noun phrase)* in unstructured texts in an unsupervised manner. This makes OpenIE highly adaptable for any domain and suitable for creation of an open knowledge graph (KG). The OpenIE methods, however, often result in generation of redundant and ambiguous information. *Canonicalization* is therefore needed to reduce redundancy and improve the quality of the resultant KG. In this work, we create a dataset for a systematic evaluation of relation canonicalization and present a quantitative analysis of existing state-of-the-art methods which has been previously missing.

Keywords: canonicalization · information extraction · knowledge graphs.

1 Introduction

Open Information Extraction (OpenIE) techniques are popularly used for the construction of knowledge graphs from raw texts [2]. However, the triples in such open KGs, e.g. Reverb [3], contain noun phrases (NPs) and relation phrases (RPs) that are not canonicalized, for example, *Obama* and *Barack Obama* refer to the same entity and *lives in* and *resident of* have the same intended meaning of the relation. Canonicalization in open KGs is the task of bringing different NPs or RPs having the same meaning to a single normalized form to improve the quality of the KG. Previous works on canonicalization in open KGs have primarily paid attention to noun phrases that represent the entities (subjects and objects) in the triples. The chief reason for this being the lack of a publicly available and large dataset against which the resulting canonicalized relation phrases could be evaluated upon. As such, only a qualitative evaluation or limited manual evaluations of the relation canonicalization has been provided so far. It is, therefore, important to evaluate the performance of existing approaches in a systematic and automated manner, so as to identify their weaknesses and further investigate the ways to improve the techniques. Towards this goal, in this work we present a large dataset comprising canonical relations and their corresponding relation phrases, which can serve as a gold standard for the evaluation of relation canonicalization methods. We describe the semi-automated process of creation of this dataset and illustrate its utility by performing the quantitative evaluation of existing state-of-the-art canonicalization approaches on it.

Related Work. Canonicalization in open KGs was discussed in detail by Galarraga et al. [4] where they showed that token overlap is an indication of similarity of NPs and RPs. They used Hierarchical Agglomerative Clustering for obtaining canonicalized clusters. Among recent works, CESI [7] used side information (including entity linking, KBP information, morphological normalization etc.) along with word vectors and KG embeddings to perform joint canonicalization of NPs and RPs. Dash et al. [1] proposed a state-of-the-art method called CUVA for canonicalization of entities and relations using variational autoencoders. It improves the canonicalization process on several fronts including entity and relation embeddings, encoding of knowledge graph structure and clustering. However, none of these methods have performed a quantitative evaluation of their performance for relation canonicalization, due to the lack of ground truth annotations for the benchmark datasets. Our work aims to fill precisely this gap.

Putri et al. [6] is one of the few works which focus on canonicalizing relations instead of entities, by aligning the relation phrases (RPs) from an open KG with the ones from Wikidata [8]. The authors show that relation alignment might be a better choice than clustering if most of the relations are likely to have equivalence in a pre-defined knowledge base. Nevertheless, the case when most of the relations of the open KG do not have their analogy in, e.g. Wikidata, is not discussed in the paper.

2 Method

To generate the dataset for relation canonicalization, our approach was to start from an existing ontological KG and derive high-quality relation phrases for its relations that can serve as golden clusters for the evaluation of the canonicalization methods. For this, we chose the NELL KG [5] (iteration 1115) which already has canonical relations. NELL was constructed in an automated way from the ClueWeb09 dataset[1] which also served as the source for other benchmark datasets often used in previous works [7,1] such as *Base, ReVerb45k* and *Ambiguous*. Figure 1 illustrates the overall steps of the dataset generation.

Selection of Representative Relations. Overall, NELL contains 832 unique relations and 2,766,048 triples. We found that not all relations were useful for the task of canonicalization. For example, very specific relations having too few representative triples in the KG would be rarely found in texts, e.g. *inverse_of_agricultural_product_coming_from_vertebrate*. On the other hand, certain relations such as *wikipedia_has_url* with a large number of triples would also be undesirable. Therefore, relations with fewer than 20 or more than 300 triples were filtered out, leaving 274 relations.

Extraction of Source sentences from NELL. The NELL dataset includes the source sentences that the triples in the KG are derived from. This serves as a useful first step for our process - for each of the relations as selected above, we consider the corresponding KG triples that the relation occurs in,

[1] http://boston.lti.cs.cmu.edu/Data/clueweb09

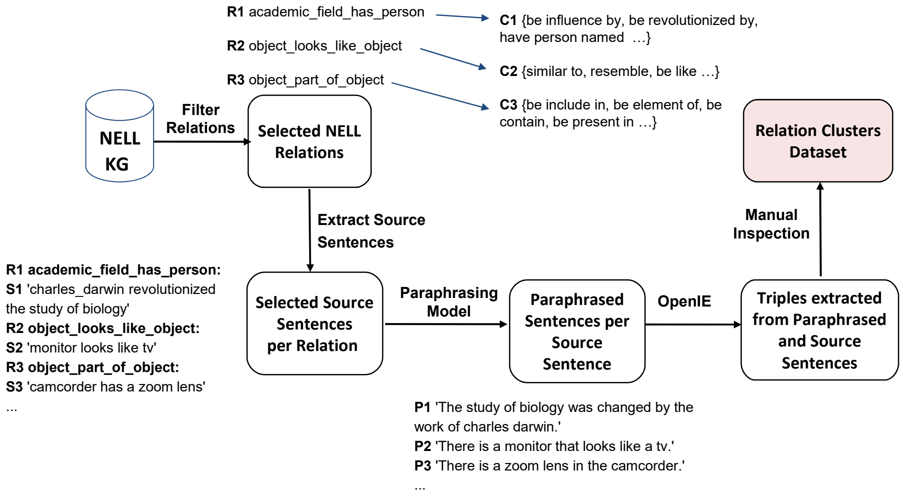
R1 academic_field_has_person
R2 object_looks_like_object
R3 object_part_of_object
C1 {be influence by, be revolutionized by, have person named ...}
C2 {similar to, resemble, be like ...}
C3 {be include in, be element of, be contain, be present in ...}
NELL KG
Filter Relations
Selected NELL Relations
Relation Clusters Dataset
Extract Source Sentences
Manual Inspection
R1 academic_field_has_person:
S1 'charles_darwin revolutionized the study of biology'
R2 object_looks_like_object:
S2 'monitor looks like tv'
R3 object_part_of_object:
S3 'camcorder has a zoom lens'
...
Selected Source Sentences per Relation
Paraphrasing Model
Paraphrased Sentences per Source Sentence
OpenIE
Triples extracted from Paraphrased and Source Sentences
P1 'The study of biology was changed by the work of charles darwin.'
P2 'There is a monitor that looks like a tv.'
P3 'There is a zoom lens in the camcorder.'
...

Fig. 1. Steps for creating an annotated dataset of relation clusters for NELL relations.

and find the source sentences for those triples. Thus, for each relation a set of source sentences is obtained that indicate the relation phrases associated with the canonical relation. Among these, the sentences having no verbs or no explicit entities were filtered out. For uniformity, the maximum number of sentences for each triple was limited to 5, leading to 73,404 sentences overall.

Derivation of Sentence Paraphrases. While the source sentences contained some phrases for the relations, we leveraged a paraphrasing model from HuggingFace [9] to obtain further relation phrases per relation[2]. The number of paraphrases for each sentence was limited to 20; thus each triple had a maximum of 120 paraphrased sentences (max 5 sentences per triple).

Extraction of Relation Phrases. In order to extract the different relation phrases in the set of paraphrased sentences, we used the Stanford OpenIE tool[3] which gave triples of the form ⟨*intuit, was eventually acquired by, mint*⟩. At this stage, the triples having no subjects or objects were filtered out and we obtained a set of relation phrases for each relation. The triples which mentioned original noun phrases in an inverse form (e.g. *yellow, is the colour of, sun* instead of *sun, has colour, yellow*) were marked as inverse and added to the set of extracted relation phrases.

Manual Inspection. The process of paraphrasing sentences automatically contributed to a wide range of possible interesting relation phrases for each relation, in many cases even better than what could have been obtained manually. However, to ensure the quality of the resulting dataset, we performed a manual cleanup to remove the noisy paraphrases. The noise for the relation varied be-

[2] https://huggingface.co/tuner007/pegasus_paraphrase

[3] https://stanfordnlp.github.io/CoreNLP/openie.html

Table 1. Results for relation canonicalization for CUVA and CESI

	Base		**Ambiguous**		**ReVerb45k**	
	Macro	*Micro*	*Macro*	*Micro*	*Macro*	*Micro*
CESI	**0.6301**	**0.5169**	**0.6284**	**0.5149**	0.6284	**0.5149**
CUVA	0.5	0.104	0.4043	0.2113	**0.6301**	0.4717

tween 10% to 50% and the took from 5 to 20 minutes depending on the relation. At this step, further nearly duplicate relations were discovered with identical RPs, e.g., *color_of_object* and *color_associated_with_visualizable_attribute*, such relations were merged.

The final dataset[4] consists of 162 canonical relations (and their inverse relations) along with their corresponding RPs, with the mean number of non-normalised RPs being 29. A few representative examples from the dataset are : **organization_acronym_has_name:** {stand for, abbreviate for, be short for} *inverse:* {full name for, be briefly know as}; **person_has_religion:** { worship, follow, believe} *inverse:* {be religion of}.

3 Evaluation and Conclusion

Quality of the Dataset. We performed manual evaluation of the generated dataset by randomly selecting 50 relations and asking two annotators to mark the corresponding RPs as correct or incorrect (1 or 0). This manual check took a couple of person hours and reported a Fleiss' Kappa agreement score of 0.80. A third annotator then independently resolved the conflicts to create the final dataset. The disagreements were mainly attributed to ambiguous or polysemous RPs that would fit well for multiple relations, as each annotator might differently imagine the necessary granularity of the dataset. The average accuracy of the RPs in the dataset after this process was 0.95.

Evaluation of Existing Methods. With our dataset serving as the ground-truth, we evaluated the relation clusters obtained from CESI and CUVA for the *Base, Ambiguous* and *ReVerb45k* datasets, with the macro and micro precision metrics [4]. The results of this evaluation, as presented in Table 1, provided some interesting insights. The scores for CESI are generally higher except for *ReVerb45k*. The reason could be that the main model used in CUVA (variational autoencoders) requires more data to learn appropriate word representations. As soon as it is provided a larger corpus, it slightly outperforms CESI. In case of *Base*, CUVA came up with about 10 relation clusters - one out of which contained over 90% of relation phrases. Such clustering results significantly affected the macro and micro precision for both *Base* and *Ambiguous* datasets.

Conclusion. In this work, we have presented a dataset for relation canonicalization that can be used for a quantitative evaluation of existing as well as

[4] https://github.com/veerlosar/rp_canonicalisation

future techniques. We hope this paves the way for further improvements in this direction. As future work, we continue to refine the proposed dataset further. In particular, we would like to avoid any bias in the dataset which could be introduced due to use of specific open-source tools such as HuggingFace and Stanford OpenIE (currently the dataset might favour the relation phrases extracted via these tools as compared to others). We plan to alleviate this issue while expanding the current dataset to include more relations and applying the pipeline to different knowledge bases. Additionally, we plan to perform a more thorough analysis for relation canonicalization and propose ways to mitigate the shortcomings of existing solutions.

References

1. Dash, S., Rossiello, G., Mihindukulasooriya, N., Bagchi, S., Gliozzo, A.: Open Knowledge Graphs Canonicalization using Variational Autoencoders. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 10379–10394 (Nov 2021)
2. Dessì, D., Osborne, F., Recupero, D.R., Buscaldi, D., Motta, E.: Generating knowledge graphs by employing Natural Language Processing and Machine Learning techniques within the scholarly domain. Future Generation Computer Systems **116**, 253–264 (2021)
3. Fader, A., Soderland, S., Etzioni, O.: Identifying Relations for Open Information Extraction. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. pp. 1535–1545 (Jul 2011)
4. Galárraga, L., Heitz, G., Murphy, K., Suchanek, F.M.: Canonicalizing open knowledge bases. In: Proceedings of the 23rd ACM International Conference on Information and Knowledge Management. pp. 1679–1688 (2014)
5. Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., Carlson, A., Dalvi, B., Gardner, M., Kisiel, B., Krishnamurthy, J., Lao, N., Mazaitis, K., Mohamed, T., Nakashole, N., Platanios, E., Ritter, A., Samadi, M., Settles, B., Wang, R., Wijaya, D., Gupta, A., Chen, X., Saparov, A., Greaves, M., Welling, J.: Never-ending learning. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15) (2015)
6. Putri, R.A., Hong, G., Myaeng, S.H.: Aligning Open IE Relations and KB Relations using a Siamese Network Based on Word Embedding. In: Proceedings of the 13th International Conference on Computational Semantics - Long Papers. pp. 142–153. Association for Computational Linguistics, Gothenburg, Sweden (May 2019)
7. Vashishth, S., Jain, P., Talukdar, P.: CESI: Canonicalizing Open Knowledge Bases using Embeddings and Side Information. In: Proceedings of the 2018 World Wide Web Conference. pp. 1317–1327 (2018)
8. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledgebase. Communications of the ACM **57**(10), 78–85 (2014)
9. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-Art Natural Language Processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45 (Oct 2020)