Survey of Evaluation Methods for Knowledge Graph Creation

Fiona Anting Tan

Institute of Data Science, National University of Singapore
tan.f@u.nus.edu

1 Introduction

In our work, we are interested to extract causal relations from text, and store these causal relations in a graph.

We could evaluate an end-to-end knowledge graph (KG) system constructed from textual inputs directly by comparing the models' suggested triples with the ground truth triples. The approach is similar to well-research fields like Named Entity Recognition and Relation Extraction. Like them, we can evaluate performance using metrics such as Precision (P), Recall (R), and F-measure.

However, merely equipped with extractive causal relations is not enough to build a meaningful KG. This is because there are multiple ways to express the same concept, and therefore, re-grouping of arguments is needed. Conversely, since context matters, it is also possible that the same expression have different meanings and should not be identified as a single meaning. In light of these issues, we have to perform argument or entity clustering. Thus, we also require evaluation methods for clustering.

This report surveys the current literature on evaluation methods for KG creation and argument or entity clustering.

2 Knowledge Graph Creation

2.1 Gold Standard Knowledge Graphs

In KG creation or completion, the aim is to capture all nodes and edges marked by a gold standard graph [10,8]. Specifically, we can source for partial gold standard graphs by asking humans to annotate a small subset of the whole KG. Alternatively, we may use other KG databases (E.g. Freebase, DBPedia, etc.) to obtain a larger gold standard graph. However, such KGs that are not annotated by humans are of a lower quality, with errors arising from target knowledge and linkages [10].

The semantic parsing challenge of the WebNLG+ Shared Task [4] involves converting English or Russian text to sets of RDF triples. The organizers created evaluation scripts that calculates P, R, and F1 scores for output triples against a gold dataset. The script searches for the optimal alignment between each suggested and true triple through all possible permutation of the hypothesis-reference pairs to be order-invariant.

Apart from P, R, F1 and Accuracy metrics, the following ranking-based metrics are popular for evaluating KGs too [1]:

- Hits@N: Indicates the number of elements in the ranking vector retrieved from the model is positioned in the top (N) locations.
- Mean Reciprocal Rank (MRR): Computes the mean of the reciprocal of elements embodied in a vector of rankings.
- Mean Rank (MR): Calculates mean rank of the correct test facts/triples embodied in a vector of rankings.

2.2 Retrospective Evaluation

We can also randomly sample some suggested completions from the automatically constructed KG, and pass these over to human judges to be marked as *correct* or *incorrect*. The evaluation metric would then be Accuracy or Precision, plus some measurement of the agreement amongst the judges [10].

[11] hired MTurk workers to rate the model performance for extracted events on 1000 news titles and for extracted causal relations on 500 pairs on a scale of 1-5. A control set, where a separate group of humans were asked to predict the Effect event given then Cause, was also rated in this manner. [6] similarly sampled 100 causal relations for review.

3 Argument/Entity Clustering

Given a gold dataset, previous research on argument or entity clustering have used the following evaluation methods [5,9,3]:

- MUC [14]: Counts the minimum number of links between mentions to be inserted or deleted when mapping a system response to a reference key set.
- B3 [2]: Precision and recall are computed from the intersection of the hypothesis and reference clusters.
- CEAF [7]: Precision and recall are computed from a maximum bipartite matching between hypothesis and reference clusters.
- NVI [12]: Information-theoretic measure that utilizes the entropy of the clusters and their mutual information. Unlike the commonly-used Variation of Information (VI) metric, normalized VI (NVI) is not sensitive to the size of the data set

In [13], they evaluate pairs against gold labels that were either *similar* or *dissimilar*, and thus could calculate F1 scores for classification.

4 Conclusion

There are multiple ways of evaluating a generated KG. These metrics help to convey the completeness and correctness of the proposed KG.

It is also important to convey the usefulness of the graph in practice. Therefore, papers introducing KGs will have to report basic summary statistics regarding the data source, number of relations, number of nodes, etc. They also analyse central concepts based on degree centrality, study the overlap of subgraphs, etc. Finally, some papers go so far as to show the graph's downstream use-cases, like in Question-Answering [6].

References

1. Abu-Salih, B.: Domain-specific knowledge graphs: A survey. J. Netw. Comput. Appl. **185**, 103076 (2021). https://doi.org/10.1016/j.jnca.2021.103076, https://doi.org/10.1016/j.jnca.2021.103076
2. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1. pp. 79–85. Association for Computational Linguistics, Montreal, Quebec, Canada (Aug 1998). https://doi.org/10.3115/980845.980859, https://aclanthology.org/P98-1012
3. Cai, J., Strube, M.: Evaluation metrics for end-to-end coreference resolution systems. In: Proceedings of the SIGDIAL 2010 Conference. pp. 28–36. Association for Computational Linguistics, Tokyo, Japan (Sep 2010), https://aclanthology.org/W10-4305
4. Castro Ferreira, T., Gardent, C., Ilinykh, N., van der Lee, C., Mille, S., Moussallem, D., Shimorina, A.: The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In: Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+). pp. 55–76. Association for Computational Linguistics, Dublin, Ireland (Virtual) (12 2020), https://aclanthology.org/2020.webnlg-1.7
5. Green, S., Andrews, N., Gormley, M.R., Dredze, M., Manning, C.D.: Entity clustering across languages. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 60–69. Association for Computational Linguistics, Montréal, Canada (Jun 2012), https://aclanthology.org/N12-1007
6. Heindorf, S., Scholten, Y., Wachsmuth, H., Ngomo, A.N., Potthast, M.: Causenet: Towards a causality graph extracted from the web. In: d'Aquin, M., Dietze, S., Hauff, C., Curry, E., Cudré-Mauroux, P. (eds.) CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020. pp. 3023–3030. ACM (2020). https://doi.org/10.1145/3340531.3412763, https://doi.org/10.1145/3340531.3412763
7. Luo, X.: On coreference resolution performance metrics. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 25–32. Association for Computational Linguistics, Vancouver, British Columbia, Canada (Oct 2005), https://aclanthology.org/H05-1004

8. Melnyk, I., Dognin, P., Das, P.: Knowledge graph generation from text. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 1610–1622. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.findings-emnlp.116
9. Moosavi, N.S., Strube, M.: Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 632–642. Association for Computational Linguistics, Berlin, Germany (Aug 2016). https://doi.org/10.18653/v1/P16-1060, https://aclanthology.org/P16-1060
10. Paulheim, H.: Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web **8**(3), 489–508 (2017). https://doi.org/10.3233/SW-160218, https://doi.org/10.3233/SW-160218
11. Radinsky, K., Davidovich, S., Markovitch, S.: Learning causality for news events prediction. In: Mille, A., Gandon, F., Misselis, J., Rabinovich, M., Staab, S. (eds.) Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012. pp. 909–918. ACM (2012). https://doi.org/10.1145/2187836.2187958, https://doi.org/10.1145/2187836.2187958
12. Reichart, R., Rappoport, A.: The NVI clustering evaluation measure. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). pp. 165–173. Association for Computational Linguistics, Boulder, Colorado (Jun 2009), https://aclanthology.org/W09-1121
13. Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., Gurevych, I.: Classification and clustering of arguments with contextualized word embeddings. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 567–578. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1054, https://aclanthology.org/P19-1054
14. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995 (1995), https://aclanthology.org/M95-1005