Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction

Martin Josifoski, Marija Šakota, Maxime Peyrard, Robert West
EPFL

Year of Publication :- 2023

Number of Citations :- 38

Introduction

Introduction

- In many applications, Large language models have demonstrated the ability to generate highly fluent and coherent textual data.
- For some Natural Language Tasks, high quality datasets are not readily available.
- For tasks where the textual input x is mapped to a structured output y, LLMs may perform poorly since, their pre-training does not gear them to produce the specified required output format.
- This paper proposes a method for generating synthetic data in the reverse direction by first sampling an output structure y and then prompting the LLM to generate a corresponding input text x.
- Closed Information Extraction is used as a task for which the dataset is generated and tested.
- Also, a model is trained on the synthetic data and compared with the baselines.

Related Works

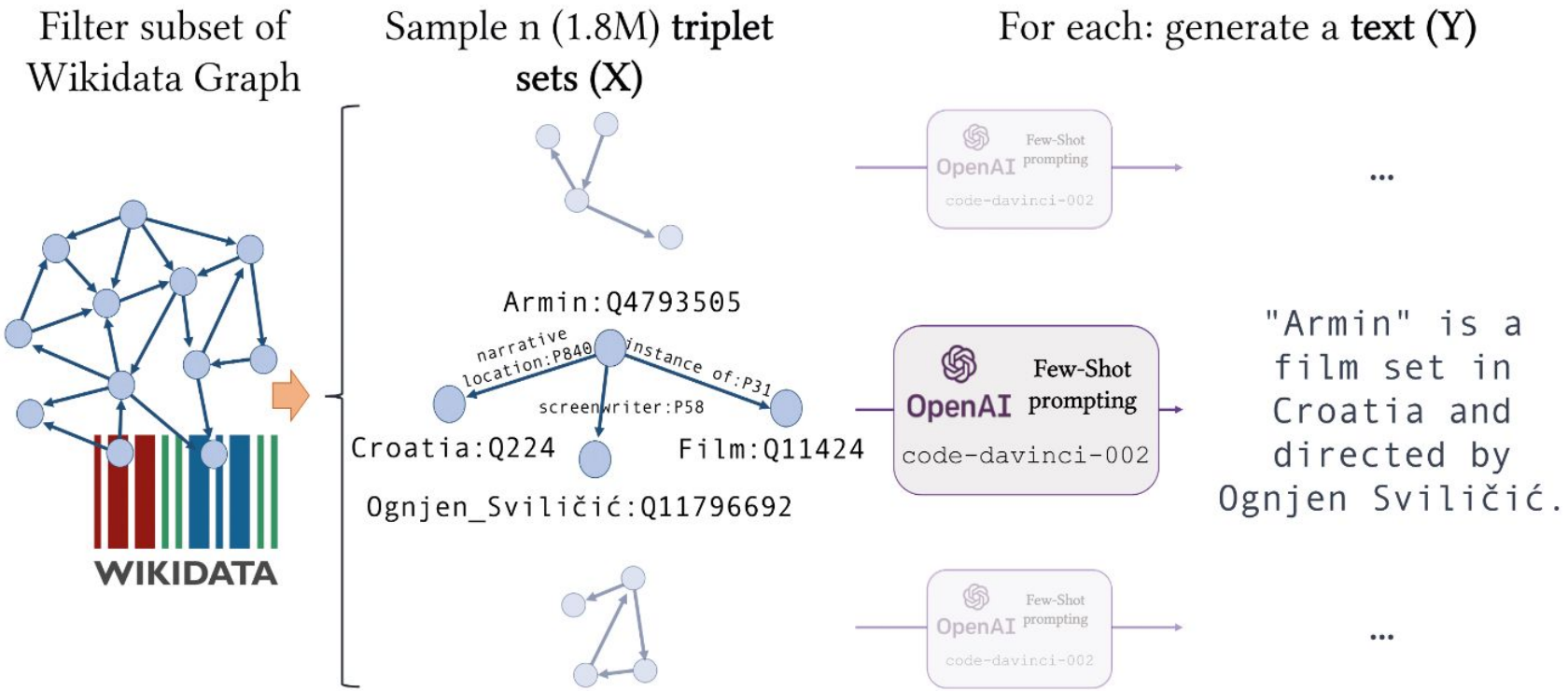
Related Works

- Meng et al. [1] and Gao et al. [2] perform the reverse task by prompting an LLM to generate comments x given a sentiment y. But, their reverse task also can be performed by an LLM.
- The largest dataset available, REBEL [3], is affected by several problems [4].
  - Noise : it is collected with a mix of heuristics, and for many data points, the target output y does not contain all the facts expressed in the input x or is (partially) incorrect.
  - Skewness : Most relations appear very rarely in the dataset, which results in models that ignore most of the information when the data is used for training.

[1] Meng, Y., Michalski, M., Huang, J., Zhang, Y., Abdelzaher, T., & Han, J. (2023, July). Tuning language models as training data generators for augmentation-enhanced few-shot learning. In International Conference on Machine Learning (pp. 24457-24477). PMLR.
[2] Gao, J., Pi, R., Lin, Y., Xu, H., Ye, J., Wu, Z., ... & Kong, L. (2022). Self-guided noise-free data generation for efficient zero-shot learning. arXiv preprint arXiv:2205.12679.
[3] Cabot, P. L. H., & Navigli, R. (2021, November). REBEL: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2370-2381).
[4] Josifoski, M., De Cao, N., Peyrard, M., Petroni, F., & West, R. (2021). GenIE: Generative information extraction. arXiv preprint arXiv:2112.08340.

Methodology

Methodology - Knowledge Graph Construction



Methodology - Knowledge Graph Construction

- From REBEL dataset, the entities that are not associated with Wikipedia pages are filtered out.
- Knowledge graph(KG) of 2.7 M entities and 888 relations with 17,655,864 edges is obtained.
- Each entity in the KG is associated with a unique English Wikipedia page title, and each relation is linked to a unique Wikidata label, which are use as their textual identifiers.
- Nodes represent entities, and edges represent relations between two entities.

Methodology - Sampling Triplet Sets

- Triplets in each set must be able to co-occur in human-written text (coherence).
- Triplet sets should be such that there is a uniform coverage of entities and relations.
- For encouraging coherence, random walk through adjacent nodes until the desired number of triplets is reached is done.
- Some entities are so central, they appear in most neighborhoods and they try to become over represented.
- After every K sampled sets, a new relation and entity distributions is formed, where the probability of sampling a specific entity or relation is inversely proportional to its frequency in the set of already sampled triplet sets S.

Methodology - Triplet-Set-to-Text Generation

- code-davinci-002 and text-davinci-003 from OpenAI's GPT 3.5 series.
- Evaluated both models in a zero-shot and a few-shot setting. In the zeroshot setting, experimented with different instructions. In the few-shot setting, instruction, the number of demonstrations, and the formatting of the demonstrations were varied.
- Experimented with different values for temperature and top-p.
- Best performing configuration selected.
- Text is generated for each triplet set.
- **Wiki-cIE Code** consists of around 1.8M training, 10K validation, and 50K test samples generated with code-davinci-002.
- **Wiki-cIE Text** consists of 10K validation and 50K test samples generated with text-davinci003 using the same triplet sets.

Experiments

Experiments

- There were experiments conducted to evaluate the benefits of training on synthetically generated data.
- The proposed model synthIE, autoregressively generates linearized sequence representation of y of the exhaustive set of facts y_{set} expressed in the input text.
- Training is done with maximizing the target sequence conditional log-likelihood with teacher forcing [5], using the cross-entropy loss, and dropout [6] and label smoothing for regularization [7].
- SynthIE is based on FLAN-T5 [8].

[5] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.
[6] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.
[7] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
[8] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70), 1-53.

Experiments

- GenIE [4] is used as the baseline model for comparison.
- Micro and Macro precision, recall and F1 are used as the evaluation metrics.
- Experiments were conducted separately for Synthetic and non synthetic data as test sets.
  - A part of the REBEL dataset that is manually annotated is used as the synthetic test set.
  - Wiki-cIE Code and Wiki-cIE Text are used as synthetic test sets.
- Also, relations are divided into different buckets according to their number of occurence in the REBEL dataset and evaluation is also done for each bucket separately.

[4] Josifoski, M., De Cao, N., Peyrard, M., Petroni, F., & West, R. (2021). GenIE: Generative information extraction. *arXiv preprint arXiv:2112.08340*.

Results

Results - Human Evaluation on REBEL dataset

- Human evaluation uncovered mistakes with already existing REBEL dataset.
  - 70% of the information from the text is not included in the “gold” set of triplets.
  - 45% of the triplets are not expressed in the input text.
- Triplets missing from the “gold” set lead to an underestimation of true precision.
- Triplets not expressed in the input text lead to overestimation of the precision and recall.
- Due to this, 360 randomly selected samples from REBEL are manually allocated. This dataset is known as $\text{REBEL}_{\text{Clean}}$.

Results

	Distant Supervision			*Synthetically Generated*					
	REBEL Clean			Wiki-cIE Text			Wiki-cIE Code		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Micro									
REBEL Gold	92.71 ± 1.73	60.68 ± 2.85	73.76 ± 2.20	–	–	–	–	–	–
GenIE T5-base	76.06 ± 3.42	51.81 ± 3.44	62.17 ± 3.01	49.10 ± 0.33	26.69 ± 0.17	34.58 ± 0.20	41.56 ± 0.49	23.94 ± 0.24	30.38 ± 0.30
SynthIE T5-base	53.02 ± 5.00	43.20 ± 3.06	48.05 ± 3.41	92.08 ± 0.17	90.75 ± 0.21	91.41 ± 0.18	79.99 ± 0.29	70.47 ± 0.30	74.93 ± 0.27
SynthIE T5-base-SC	59.97 ± 4.34	30.54 ± 2.14	40.76 ± 2.57	92.79 ± 0.12	90.50 ± 0.10	91.63 ± 0.10	81.58 ± 0.15	69.48 ± 0.29	75.05 ± 0.19
SynthIE T5-large	68.25 ± 4.91	54.37 ± 3.08	61.26 ± 3.07	93.38 ± 0.11	92.69 ± 0.19	93.04 ± 0.13	82.60 ± 0.19	73.15 ± 0.29	77.59 ± 0.24
Macro									
REBEL Gold	51.21 ± 5.03	41.02 ± 4.69	43.76 ± 4.62	–	–	–	–	–	–
GenIE T5-base	39.36 ± 4.68	31.46 ± 4.24	33.33 ± 4.07	29.82 ± 0.67	11.14 ± 0.15	13.94 ± 0.17	25.78 ± 0.85	9.81 ± 0.10	12.12 ± 0.12
SynthIE T5-base	35.57 ± 4.82	34.05 ± 4.47	33.13 ± 4.44	94.10 ± 0.15	92.42 ± 0.17	93.05 ± 0.11	83.76 ± 0.36	74.05 ± 0.45	77.91 ± 0.42
SynthIE T5-base-SC	20.07 ± 3.26	12.82 ± 2.65	14.65 ± 2.70	94.35 ± 0.19	92.39 ± 0.20	93.15 ± 0.15	84.32 ± 0.32	73.57 ± 0.41	77.88 ± 0.34
SynthIE T5-large	54.11 ± 5.26	52.01 ± 4.64	51.04 ± 4.76	95.27 ± 0.22	94.95 ± 0.13	94.99 ± 0.12	86.43 ± 0.25	78.78 ± 0.27	81.95 ± 0.22

Results

- The original annotations of REBEL score a Micro F1 73.8 and Macro F1 43.76 with the manually annotated test data. This indicates that REBEL dataset has an unsatisfactory performance for a dataset which is used to evaluate model performance.
- Observations from the experiments done with the $\text{REBEL}_{\text{clean}}$ as the test dataset.
  - SynthIE T5-large outperforms REBEL Gold in Macro F1.
  - SynthIE T5-base is on par with GenIE T5-base in Macro F1.

Results

- $\text{REBEL}_{\text{clean}}$ dataset also exhibits some issues similar to REBEL.
  - A high imbalance in terms of the relation occurrence counts.
  - Text often containing information for entities that cannot be resolved.
- These findings emphasize the importance of the proposed Wiki-cIE Text as a reliable evaluation dataset for the cIE task.
- Also, it is seen that the $\text{Gen-T5}_{\text{base}}$ performance degrades by more than 50% with Synthetic datasets. In REBEL, the model performs well on a few relations that occur more frequently and badly on the rest. In synthetic data, as the relations are balanced, $\text{Gen-T5}_{\text{base}}$ performance degrades.

Results

- SynthIE T5-base, which is trained on data synthetically generated by the proposed methodology, and differs from GenIE T5-base only in terms of the training data, achieves a 91.4 micro, and an even higher 93.1 macro-F1 score on Wiki-cIE Text.
- Also, when analyzing the performance by relation frequency, it is found that for Relations that occur less than 32, the performance of GenIE-$T5_{base}$ is close to 0.

Conclusion

- A large dataset is synthetically generated which can address the issues faced by some of the commonly used datasets for cIE.
- Some of the issues faced by the model trained in the REBEL dataset is addressed.
- The model trained on synthetic data outperformed the models trained on the REBEL data.

References

1. Meng, Y., Michalski, M., Huang, J., Zhang, Y., Abdelzaher, T., & Han, J. (2023, July). Tuning language models as training data generators for augmentation-enhanced few-shot learning. In International Conference on Machine Learning (pp. 24457-24477). PMLR.
2. Gao, J., Pi, R., Lin, Y., Xu, H., Ye, J., Wu, Z., ... & Kong, L. (2022). Self-guided noise-free data generation for efficient zero-shot learning. arXiv preprint arXiv:2205.12679.
3. Cabot, P. L. H., & Navigli, R. (2021, November). REBEL: Relation extraction by end-to-end language generation. In Findings of the Association for Computational Linguistics: EMNLP 2021 (pp. 2370-2381).
4. Josifoski, M., De Cao, N., Peyrard, M., Petroni, F., & West, R. (2021). GenIE: Generative information extraction. arXiv preprint arXiv:2112.08340.
5. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.
6. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.
7. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
8. Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., ... & Wei, J. (2024). Scaling instruction-finetuned language models. Journal of Machine Learning Research, 25(70), 1-53.
9. Josifoski, M., De Cao, N., Peyrard, M., Petroni, F., & West, R. (2021). GenIE: Generative information extraction. arXiv preprint arXiv:2112.08340.