Extraction and Generation Tasks with Knowledge-aware Text-to-Text Transfer Transformer

Mohammad Golam Sohrab[1] Makoto Miwa[1,2]
[1]Artificial Intelligence Research Center (AIRC)
National Institute of Advanced Industrial Science and Technology (AIST)
[2]Toyota Technological Institute, Japan
sohrab.mohammad@aist.go.jp, makoto-miwa@toyota-ti.ac.jp

Abstract

We introduce knowledge-aware transfer learning with a text-to-text transfer transformer (KAT5) by leveraging a text-to-text transfer transformer (T5) in the Wikipedia domain. In standard transfer learning like T5, a model is first pre-trained on an unsupervised data task with a language model objective before fine-tuning it on a downstream task. In this work, we align large-scale alignments between Wikipedia abstract and Wikidata triples to facilitate our pre-training KAT5 model. Experiment result shows that KAT5 can match or outperform several downstream tasks, including question answering, entity and relation extraction, summarization and machine translation.

1 Introduction

In this work, to better capture the awareness of knowledge in language modeling pre-training, we present a knowledge-aware text-to-text transfer transformer that packs more information into the T5 model [1], which we call KAT5. During transfer learning a model is first pre-trained on a large-scale unsupervised data task and the most successful approaches have been variants of masked language models (MLMs), which are denoising autoencoders that are trained to reconstruct text by masking out a random subset of the input sequence.

Integrating knowledge like entity or coreference information during transfer learning in NLP is not a common fashion as it needs to label a large-scale dataset. Such large-scale label dataset is not available, therefore, it is common to pre-train the entire model using data-rich unsupervised learning on unlabeled data. Our baseline model, T5 investigates different objective tasks, including masked language model (MLM), random span, and deshuffling, where the model is limited to exploring integrating knowledge during pre-training. Here, we push the limits of this model by grafting knowledge like entity and co-reference information by mapping Wikipedia and Wikidata during pre-training. We perform large-scale alignments between Wikipedia abstract and Wikidata triples to facilitate our pre-training KAT5 model and further research on integrating knowledge into large-scale pre-training. We show that initialization with knowledge-aware pre-training is effective for various downstream tasks. We fine-tune and evaluate the KAT5 model in joint entity-relation extraction and generation tasks− question answering, abstractive summarization, and machine translation. We compare its performance with several recent state-of-the-art models.

The KAT5 implementation is based on Huggingface transformers. This work is a short version of our previously published KAT5 [2] model, where the question-answering task on two datasets are additionally evaluated.

2 Model Architecture

We consider the text-to-text transfer transformer – T5 [1] as a baseline encoder-decoder architecture close to the original architecture of transformer [3].

2.1 KAT5: Knowledge-Aware Text-To-Text Transfer Transformer

As an unsupervised objective during pre-training, a model needs an objective function that does not require labels but teaches the model generalizable knowledge and will be useful to transfer that knowledge into downstream tasks. Apart from casual language modeling objective for pre-training, recently denoising a.k.a. masked language modeling (MLM) shows better performance and become

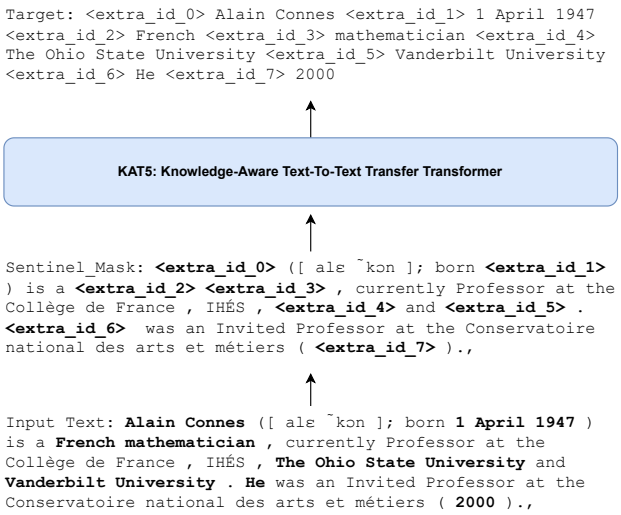

Figure1 Pre-training tasks of KAT5

a standard unsupervised learning objective in many natural language processing (NLP) tasks. In the MLM objective, the model is trained to predict missing or corrupted tokens by adding <MASK> in the input sequence. Inspired by BERT' s MLM objective, T5 follows a random span masking objective to corrupt 15% of tokens in the input sequence where all consecutive spans of dropped-out tokens are replaced by a single sentinel token, a.k.a. unique mask tokens. We adopt the T5 masking strategies and design an objective that randomly samples and then drops out 15% of entity and coreference related spans in the input sequence using 100 sentinel tokens. Each sentinel token represents a unique mask token starting as <extra_id_0>, <extra_id_1>, ⋯ , <extra_id_99> for a given input sequence.

Fig. 1 shows a knowledge-aware task of KAT5. In this figure, the bold text in the input sequence represents entities where the pronoun He indicates the coreference of Alain connes. During sentinel masking in KAT5, unique mask tokens are used to corrupt the input text by replacing the entity and coreference spans. Finally, the output sequence consists of the dropped-out entity and coreference spans, delimited by the sentinel tokens used to replace them in the input.

2.2 Pre-training Data Creation

Another key contribution of this paper is to automatically create data for pre-training the KAT5 model. The pre-training data set is a crucial component of the transfer learning pipeline. During pre-training, the model needs a large amount of data that teaches the model generalizable knowledge. The T5 model used the Colossal Clean Crawled Corpus (C4) dataset for pre-training by downloading about 750 GB of text extracted from the Web. In contrast, our KAT5 model is based on integrating knowledge like entity and co-reference information during pre-training. One possible way is to create such a knowledge-aware pre-training dataset by aligning Wikipedia abstract and Wikipedia hyperlinks with Wikidata entities. We aligned the Wikipedia abstract and Wikidata entities to pre-train the KAT5 model to shed light on this challenging task. We create the knowledge-aware pre-training dataset by adopting the T-REx implementation [1)]. In this implementation, we integrate entity or mention type predictors using the spaCy [2)] model to predict all the span types of Wikipedia links. For space limitation, we refer the readers to [2] for more details of pre-training data creation.

3 Experimental Settings

3.1 Datasets

Several datasets for different downstream tasks, including SQuAD 1.1 [4] and SQuAD 2.0 [4] datasets for question answering, CoNLL04 [5], ADE [6], and NYT [7] datasets for joint entity-relation extraction tasks, XSum [8] and CNN/DailyMail (CNNDM) datasets for summurization tasks, and the WMT shared-task datasets from Hugging Face [3),4)] are used to evaluate our KAT5 model. We refer to Appendix A for more details about the datasets.

3.2 KAT5 Pre-training and Fine-tuning

To pre-train the KAT5 model, we initialize the model with the T5-base checkpoint [5)] and continue pre-training using the knowledge-aware span denoising objective of T5 on the training split of our dataset that was explained in Section 2.2. During KAT5 fine-tuning on downstream datasets, Like T5 [1], we treat every text processing problem as a test-to-text problem, i.e. giving text as input to the KAT5 model and producing new text as output. We consider two learning settings - (1) Single-task learning:

1) https://github.com/hadyelsahar/RE-NLG-Dataset
2) https://spacy.io
3) https://huggingface.co/datasets/wmt14
4) https://huggingface.co/datasets/wmt16
5) https://huggingface.co/google-t5/t5-base

Table1 Performance comparison on the SQuAD 1.1 and SQuAD 2.0 datasets.

Model	SQuAD 1.1		SQuAD 2.0	
	EM	F1	EM	F1
BNA [4]	68.0	77.3	59.8	62.6
DocQA [4]	71.1	81.0	61.9	64.8
DocQA + ELMo [4]	78.6	85.8	65.1	67.6
T5 [1]	80.8	-	-	-
KAT5	**81.5**	**88.4**	**78.0**	**81.5**

a single model on a single dataset is learned initializing from KAT5 checkpoint. (2) Multi-task learning: Since the model is based on our direct baseline T5 model, therefore, our KAT5 model naturally allows us to train a single model on multiple datasets that can cover many structured prediction tasks. Learning parameters for pre-training and fine-tuning are discussed in Appendix B.

4 Results

We show that our Knowledge-aware T5 (KAT5) can effectively solve the structure prediction tasks that match or exceed the previous state of the art on multiple datasets. To evaluate our model, we adopt TANL [9] evaluation script for joint entity-relation extraction tasks and Hugging Face Transformers evaluation script for question-answering, summarization and translation tasks.

4.1 Question Answering

Table 1 shows the question answering performance comparison of KAT5 model over the SQuAD 1.1 and SquAD 2.0 datasets. KAT5 outprforms our direct baseline T5 model and shows a significant improvements over the other models in this table.

4.2 Joint Entity-Relation Extraction

We tackle the joint entity-relation as a generation task where the model output of KAT5 is a triplet that is present in the input text. With the single-task setup in Table 2, the KAT5 outperforms over the TANL which is our direct baseline since TANL framework is initialized with T5 and used the same model parameters. We obtain a +0.6/+1.7/+0.2 and -0.6/+1.0/+0.1 improvement using F1 score in the CONLL4/ADE/NYT datasets for entity and relation extraction tasks respectively. TANL, needs 200 epochs to achieve the stated results in Table 2 where we fine-tune on top of KAT5 for 10 epochs. In contrast to our baseline approaches, KAT5 shows a better performance over the SpERT [10], but shows a little drop in comparison to the Rebel [11] which is a task specific model.

4.3 Summarization

Table 3 shows the abstractive performance comparison of KAT5 over the XSum dataset. Both the single- and multi-task settings, the KAT5 outperforms the baseline T5 model. The model also outperforms BART and the recent non-autoregressive BERT-NAR-BERT model. Table 4 shows the performance comparison of KAT5 over the CNNDM dataset. The model shows an improvement over the baseline model but shows a little drop in comparison to the BART model.

4.4 Machine Translation

Results of machine translation (MT) experiments are summarized in Table 5. The KAT5 model outperforms the baseline T5 that obtains a +0.36/+3.01 improvement using BLEU score in the EN-DE/EN-RO WMT datasets respectively. In comparison to the non-autoregressive benchmark, the KAT5 model outperforms all formats of the BERT-NAR-BERT models.

5 Discussion

We present Knowledge-aware T5 (KAT5), a novel, simple, and easy-to-implement S2S model by leveraging T5 checkpoint during pre-training. We demonstrate strong performances of joint entity-relation extraction in three datasets (ADE, CONLL04, and NYT), XSum and CNNDM in summarization tasks, and English (EN) → German (DE) and English→Romanian (RO) in machine translation. KAT5 is a budget training approach since it needs 10 epochs that can achieve similar or somewhat better performance over the each CONLL04, NYT, and ADE datasets where TANL set 200 epochs to achieve the reported score in Table 2 for all the entity-relation extraction datasets.

6 Related Work

T5 [1] - the basic idea underlying this work is to treat every text processing problem as a "text-to-text" problem, i.e. taking text as input and producing new text as output. The model achieves state-of-the-art results on many benchmarks covering summarization, question answering,

Table2 Performance comparison on the CONLL04, ADE, and NYT datasets. Bold and underlined denotes the best and second-best results within KAT5 and Baseline Models.

Model	Params	CONLL04		ADE		NYT	
		Entity	Relation	Entity	Relation	Entity	Relation
SpERT [10]	110M	88.9	71.5	89.3	78.8	-	-
REBEL_pretraining [11]	460M	-	75.4	-	82.2	-	92.0
- Baseline Model -							
TANL + Single-task [9]	220M	89.4	71.4	90.2	80.6	94.9	90.8
KAT5 + Single-task	220M	90.0	69.8	**91.9**	81.6	**95.1**	**90.9**

Table3 Performance comparison on the XSum dataset. R-1/2/L stands for ROUGE-1/2/L.

Model	XSum		
	R-1	R-2	R-L
Transformer [3]	30.7	10.8	24.5
ELMER-Soft [12]	38.3	14.2	29.9
BART [13]	38.8	16.2	30.6
BERT2BERT [14]	37.5	15.2	30.1
BnB + additional pre-training [15]	36.1	13.4	30.0
- Baseline Model -			
T5 + fine-tuning + Single-task	39.7	16.5	31.9
KAT5 + Single-task	39.9	16.7	32.1
KAT5 + Multi-task	**40.2**	**17.0**	**32.2**

Table4 Performance comparison on the CNN/DailyMail (CNNDM) dataset. Bold and underlined scores denote the best and second-best results within KAT5 and Baseline Models.

Model	CNNDM		
	R-1	R-2	R-L
BERTSUMABS [16]	41.72	19.39	38.76
BERTSUMEXTABS [16]	42.13	19.60	39.18
ROBERTASHARE [14]	40.31	18.91	37.62
BART [13]	44.16	21.28	40.90
- Baseline Model -			
T5 [1]	-	19.24	-
KAT5 + Single-task	**43.51**	**20.64**	**40.66**
KAT5 + Multi-task	43.44	20.28	40.56

text classification, and more. We adopt this approach as our direct baseline by grafting knowledge like entity and coreference information during pre-training.

TANL [9] - a framework to solve several structure predictions in a unified way, with a common architecture and without the need for task-specific modules. This is our baseline approach for joint entity and relation extraction tasks as the model initializes from the T5-base model like our approach during pre-training.

Table5 Machine translation experiment results in BLEU scores.

Model	EN - DE	EN - RO
Transformer [3]	27.30	21.53
BERT2BERT + mBERT [14]	25.80	23.24
BnB + mBERT + distilled [15]	27.49	18.94
- Baseline Model -		
T5 [1]	27.65	26.98
KAT5	**28.01**	**29.99**

REBEL [11] - a sequence-to-sequence (S2S) model based on BART-large that performs end-to-end relation extraction for more than 200 different relation types and show that how relation extraction can be simplified by expressing triplets as a sequence of text.

BART-NAR-BERT (BnB) [15] – a pre-trained non-autoregressive S2S model, which employs BERT as the backbone for the encoder and decoder for natural language understanding and generation tasks. The model outperformed several SOTA models in non-autoregressive benchmark and has shown comparable performance in autoregressive models. Since the model follows a S2S manner, we also compare our model over the generative tasks.

7 Conclusion

This paper introduces an efficient Knowledge-aware T5 (KAT5) S2S method with encoders and decoders that integrates entities and their coreferences as knowledge during pre-training. To introduce such knowledge-aware approach, we perform large-scale alignments between Wikipedia abstract and Wikidata triples to facilitate our pre-training KAT5 model by leveraging T5 model. Experiment results show that the proposed model outperforms baselines in most of the downstream tasks. In the future, we plan to extend our KAT5 model into a larger parameter model with more knowledge-aware data.

Acknowledgement

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. **Journal of Machine Learning Research**, Vol. 21, No. 140, pp. 1–67, 2020.

[2] Mohammad Golam Sohrab and Makoto Miwa. Kat5: Knowledge-aware transfer learning with a text-to-text transfer transformer. In Albert Bifet, Tomas Krilavičius, Ioanna Miliou, and Slawomir Nowaczyk, editors, **Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track**, pp. 157–173, Cham, 2024. Springer Nature Switzerland.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. **Advances in neural information processing systems**, Vol. 30, , 2017.

[4] Pranav Rajpurkar, Jian Zhang, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In **ACL 2018**, 2018.

[5] Dan Roth and Wen-tau Yih. A linear programming formulation for global inference in natural language tasks. In **Proceedings of the Eighth Conference on CoNLL-2004 at HLT-NAACL 2004**, pp. 1–8, Boston, Massachusetts, USA, May 6 - May 7 2004. Association for Computational Linguistics.

[6] Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. **Journal of Biomedical Informatics**, Vol. 45, No. 5, pp. 885–892, 2012.

[7] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In Iryna Gurevych and Yusuke Miyao, editors, **Proceedings of the 56th Annual Meeting of the ACL (Volume 1: Long Papers)**, pp. 506–514, Melbourne, Australia, July 2018.

[8] Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on EMNLP**, pp. 1797–1807, Brussels, Belgium, October-November 2018. ACL.

[9] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In **9th International Conference on Learning Representations, ICLR 2021**, 2021.

[10] Markus Eberts and Adrian Ulges. Span-based joint entity and relation extraction with transformer pre-training. **CoRR**, 2019.

[11] Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Findings of the Association for Computational Linguistics: EMNLP 2021**, pp. 2370–2381. ACL.

[12] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. ELMER: A non-autoregressive pre-trained language model for efficient and effective text generation. In **Proceedings of the 2022 Conference on EMNLP**, pp. 1044–1058, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[13] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 7871–7880. ACL, July 2020.

[14] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. Leveraging pre-trained checkpoints for sequence generation tasks. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 264–280, 2020.

[15] Mohammad Golam Sohrab, Masaki Asada, Matīss Rikters, and Makoto Miwa. Bert-nar-bert: A non-autoregressive pre-trained sequence-to-sequence model leveraging bert checkpoints. **IEEE Access**, Vol. 12, pp. 23–33, 2024.

[16] Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, **Proceedings of the 2019 Conference on EMNLP-IJCNLP**, pp. 3730–3740, Hong Kong, China, November 2019. ACL.

[17] Pankaj Gupta, Hinrich Schütze, and Bernt Andrassy. Table filling multi-task recurrent neural network for joint entity and relation extraction. In Yuji Matsumoto and Rashmi Prasad, editors, **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics**, pp. 2537–2547.

[18] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. **CoRR**, Vol. abs/2101.05779, , 2021.

[19] Bowen Yu, Zhenyu Zhang, Jianlin Su, Yubin Wang, Tingwen Liu, Bin Wang, and Sujian Li. Joint extraction of entities and relations based on a novel decomposition strategy. **CoRR**, Vol. abs/1909.04273, , 2019.

[20] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In **Text Summarization Branches Out**, pp. 74–81, Barcelona, Spain, July 2004. ACL.

A Dataset Details

A.1 Question Answering Dataset

SQuAD 1.1 The Stanford Question Answering Dataset (SQuAD 1.1) [4] is a large reading comprehension dataset on Wikipedia articles.

SQuAD 2.0 The SQuAD 2.0 [4] is a new dataset that combines answerable questions from the previous version of SQuAD (SQuAD 1.1) with 53,775 new, unanswerable questions about the same paragraphs.

A.2 Joint Entity-Relation Extraction Dataset

CONLL04 The CoNLL04 [5] dataset consists of sentences extracted from news articles - with four entity types `location`, `organization`, `person`, and `other`, and five relation types `(work for`, `kill`, `organization based in`, `live in`, and `located in`. We use the 922/231/288 sentences in the train/validation/test set based on the split by Gupta [17].

ADE The ADE [6] dataset consists of 4,272 sentences extracted from medical reports– with two `drug` and `disease` entity types and a single relation type `effect`. This dataset has sentences with nested entities. We follow the same settings as TANL [18], conduct a 10-fold cross-validation, and report the average macro-F1 results across all ten splits.

NYT The NYT dataset [7] is based on the New York Times corpus, where we use the preprocessed version of Yu [19]. It consists of three entity types `location`, `organization`, `person` and 24 relation types (such as place of birth, nationality, company, etc.). It consists of 56,195/5000/5000 sentences in the training/validation/test set.

A.3 Summarization Dataset

XSum [8] Abstractive text summarization aims to produce a short version of a document while preserving its salient information content. We evaluate the models based on the BBC extreme [8] (XSum) dataset. This is a news summarization dataset containing 227K news articles and single-sentence summary pairs. We load the XSum datasets from Huggingface [6] The evaluation metric is ROUGE [20], including ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L). We adopted the Google Research re-implementation of ROUGE[7].

CNNDM The CNN/DailyMail (CNNDM) dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. We load the CNNDM datasets from Hugging Face datasets [8] that supports both extractive and abstractive summarization.

A.4 Machine Translation Dataset

We evaluate our models using two popular benchmark data sets from the WMT shared tasks on news translation - English (EN) → German (DE) data from WMT 2014 and English→Romanian (RO) data from WMT 2016. We load the WMT datasets from Hugging Face datasets[9),10)] and use them directly to train the models without filtering. We evaluate the performance by computing BLEU.

B Learning Parameters

We use a learning rate of 0.001, a linear warm-up of 5k steps, a gradient accumulation of 2 steps, and a maximum sequence length of 512 tokens. The KAT5 model is trained on 1.3B tokens, where we employ a batch size of 65,536 tokens with a maximum step of 200K steps. The original T5 model was trained on 34B tokens over the C4 corpus, which was 26 times larger than our additional pre-training dataset. The KAT5 model is optimized end-to-end using an Adafactor optimizer with a corrupted knowledge-aware span ratio of 15%.

We fine-tune on top of KAT5 for a maximum of 10 epochs in all our downstream tasks. In the multi-task settings of summarization tasks, we add the dataset name followed by the task separator is used (for example, “xsum summarize :” for XSum dataset and “summarize :” for CNNDM dataset) as a prefix to each input sentence.

6) https://huggingface.co/datasets/EdinburghNLP/xsum

7) https://github.com/google-research/google-research/tree/master/rouge

8) https://huggingface.co/datasets/cnn_dailymail

9) https://huggingface.co/datasets/wmt14

10) https://huggingface.co/datasets/wmt16