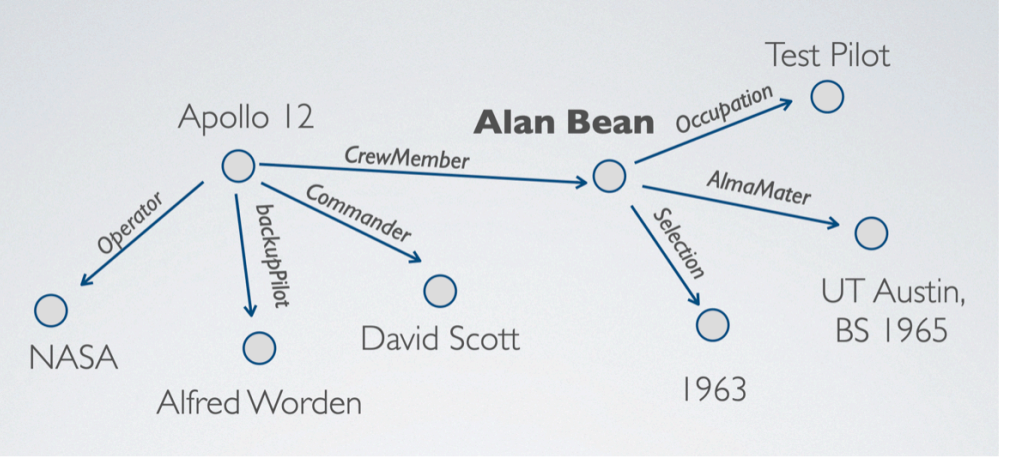
Text and Knowledge Graphs

Similarity and Generation

Claire Gardent

CNRS / LORIA, Nancy

Texts and Knowledge Graphs



Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred. Worden and was commanded by David Scott

Motivations

- Cross-modal Graph-Text retrieval
- Evaluation
  - KG-to-Text generation
  - *Does the text generated convey all and only the information represented by the input knowledge graph?*
- KG-to-Text Generation
  - *Can we use a KG-Text similarity metrics to guide generation ?*

Outline

A Joint Encoder for KGs and English texts

- Retrieval
- Reference less evaluation of KG-to-Text Generation

A Fine-Grained Similarity Metrics for Text and Knowledge Graphs

- Multilingual
- Regression Model pre-trained on NLI (Natural Language Inference) data and fine-tuned on KG-Text pairs
- Fine-Grained
  - Recall: how much does the generated text convey the content of the input graph ?
  - Precision: how much of the generated text is factually consistent with the input graph?

DPO-guided KG-to-Text Generation

- Create preference data using a KG-Text similarity metrics
- Fine tune an instruction tuned decoder on this preference data using Direct Preference Optimisation (DPO)

EreDat: A Similarity Metric for English Texts and Knowledge Graphs

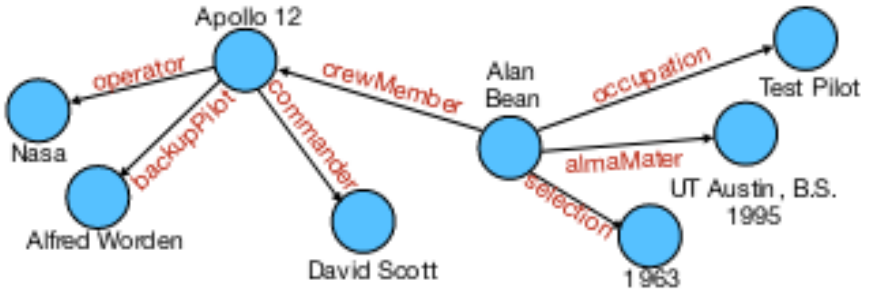
T. Le Scao and C. Gardent. Joint Representations of Text and Knowledge Graphs for Retrieval and Evaluation In Findings of IJCNLP-AACL 2023

Semantic Similarity

Between

- words
  (Mikolov et al., 2013; Devlin et al., 2019)
- sentences
  (Reimers and Gurevych, 2019: Chen et al., 2020; Humeau et al., 2019)
- Knowledge Base entities and relations
  (Bordes et al., 2013; Yang et al., 2014; Trouillon et al., 2016; Dettmers et al., 2018; Schlichtkrull et al., 2018)
- Images and text
  (Radford et al. 2021)

Joint Encoder for English text and RDF Graphs



Alan Bean graduated from UT Austin in 1955 with a Bachelor of Science degree. He was hired by NASA in 1963 and served as a test pilot. Apollo 12's backup pilot was Alfred Worden and was commanded by David Scot

Challenge: Lack of parallel data

Silver Data for training

TeKGen. 6M Wikidata graphs heuristically aligned with Wikipedia sentences.

KELM. 15M (Wikidata graph, text) pairs where the text is automatically generated from the graph.

TREx. 11M Wikidata triples heuristically aligned with 6 million Wikipedia sentences.

	# (t,g)	# P	# E
TeKGen	6,310,061	1041	3,939,696
TREx	6,000,336	675	3,188,309
KELM	15,616,551	261405	5,073,603
WebNLG-DB	13,212	372	3210
WebNLG-WD	10,384	188	2783
WikiChunks	30,000	468	20,318

Test Data

WebNLG-DB 13K parallel (graph,text) pairs where the texts were crowdsourced to match the input graph and the graph is extracted from the DBpedia KB.

WebNLG-WD 10K parallel (graph,text) pairs where the text is a text from WebNLG-DB and the corresponding DBPedia graph has been mapped to Wikidata.

	# (t,g)	# P	# E
TEKGEN	6,310,061	1041	3,939,696
TREX	6,000,336	675	3,188,309
KELM	15,616,551	261405	5,073,603
WEBNLG-DB	13,212	372	3210
WEBNLG-WD	10,384	188	2783
WIKICHUNKS	30,000	468	20,318

WikiChunks 7.3M graph-text pairs where the text is a 100-word *passage* from a Wikipedia dump and the graphs are matching Wikidata graphs.

Model

Bi-encoder

- Mean-pooling to create fixed-sized embeddings for KGs and texts
- Contrastive loss with in-batch negatives

$$l = -\sum_{i \in I} \log \left(\frac{\exp(sim(text_i, kg_i))}{\sum_{j \in J} \exp(sim(text_i, kg_j))} \right)$$

  - Maximise the similarity of matching KG-Text pairs
  - Multi-class classification problem: each text must be matched to its matching KG. We compute the pairwise similarities between each (graph, text) pair in the batch and apply a softmax on the KG axis.

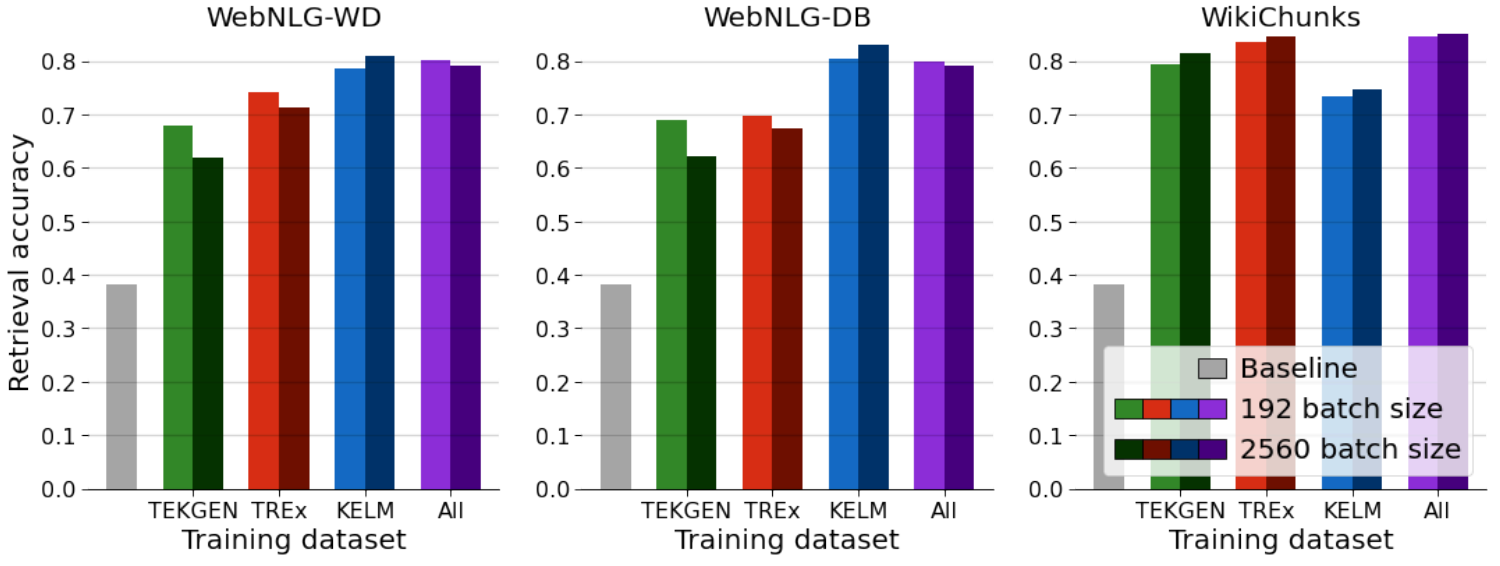
Baseline

all-mpnet-base-v2

- A state-of-the-art sentence embedding model
- optimised to assess semantic similarity between texts
- used to initialise our bi-encoder

Retrieval Accuracy

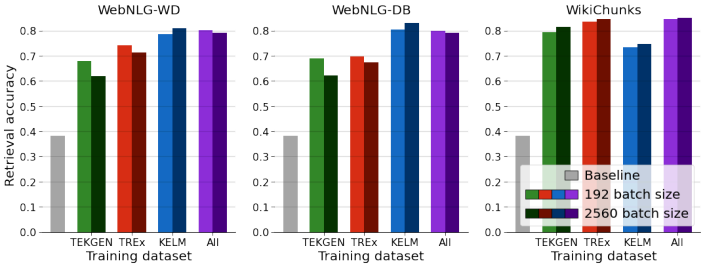
Given the embedding of a graph, how well can we identify the most similar text in the corpus ?

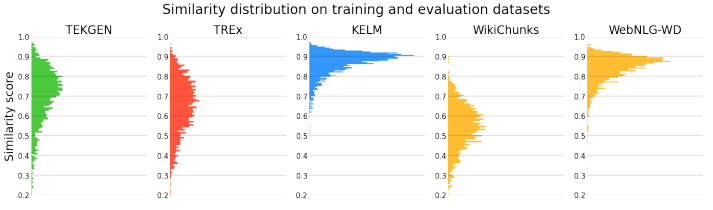


Retrieval Accuracy

Impact of Training Data

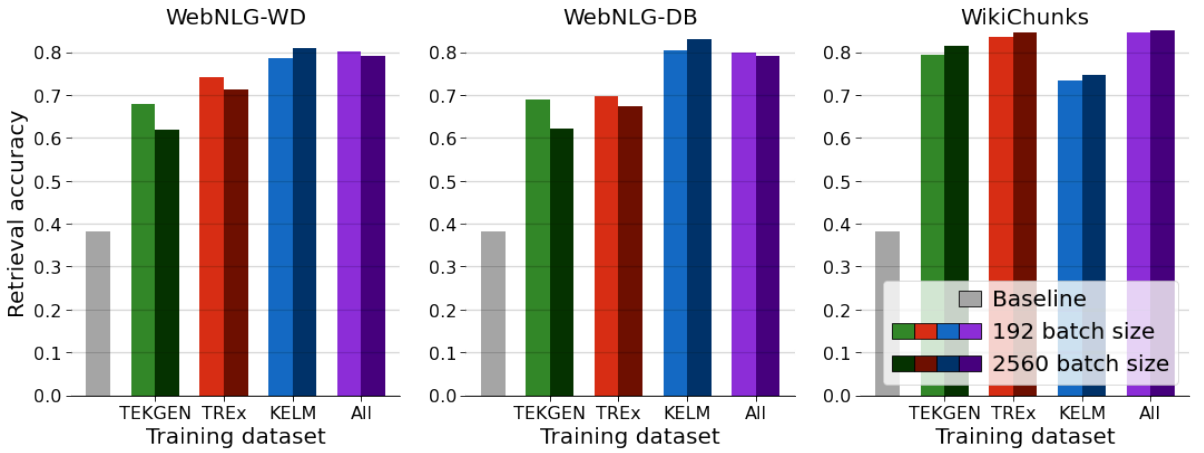
- Large improvement over the baseline
- Accuracy varies with the training data used
- Better aligned data results in better retrieval accuracy





Retrieval Accuracy

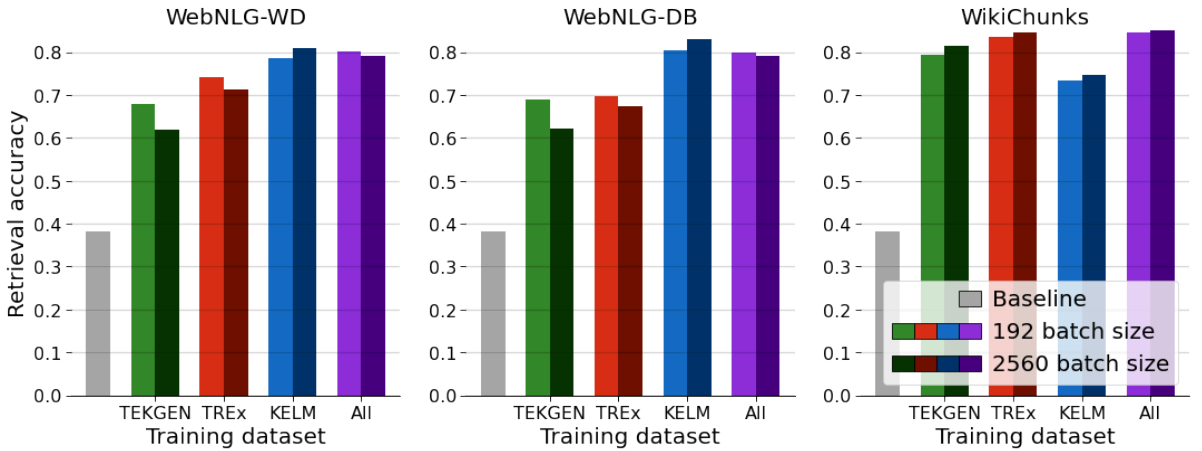
Generalising to other KB



Trained on Wikidata: Similar Results when Testing on DBPedia

Retrieval Accuracy

Testing on parallel vs. Noisy data



Better results on Wikichunks as it is more similar (noisy alignment) to the training data

Retrieval Accuracy

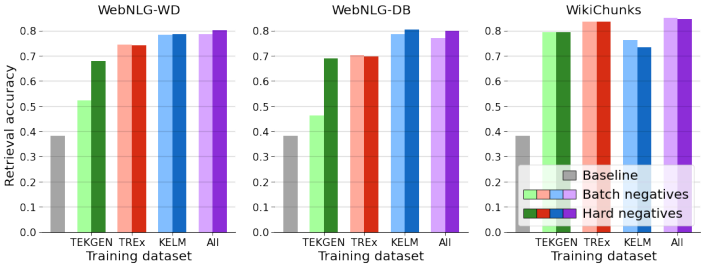
Hard vs. In Batch Negatives

Hard negatives

- The graph is corrupted by replacing a subject, object or predicate at random by another resource in the data set.

Hard negatives mostly help

- when retrieving on parallel data (WebNLG) i.e., when small graph-text mismatches strongly impact accuracy.
- when the training data is most noisy (TekGen)



Hard negatives are most helpful when the training data is noisier than the evaluation data.

Evaluation Metric for KG-to-Text Models

We further improve the model by

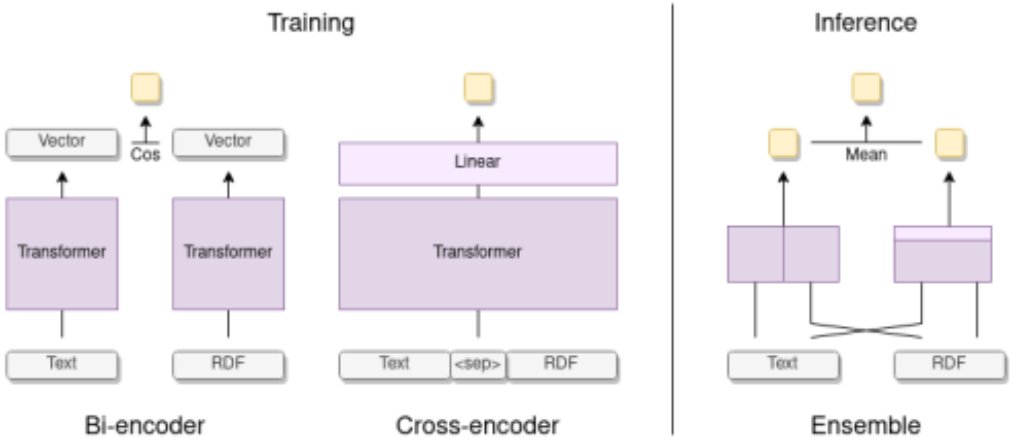
- fine-tuning on human judgments of KG-text similarity
- ensembling a bi and a cross-encoder
- adding inverted negatives

Fine-tuning on human judgments of KG-text similarity

- WebNLG 2017
- 2,230 generated texts (10 models) annotated with human judgments of *semantic adequacy*

  Does the text correctly represent the meaning in the data?

Bi- and Cross-Encoder



Bi-encoder: Text and graphs are encoded separately

Cross-encoder: One model instance attends to both text and graphs simultaneously

Ensembling: The mean of the bi- and cross-encoder scores

Inverted Negatives

Triple

(***André the Giant***, larger than, Samuel Beckett)

Inverted Triple

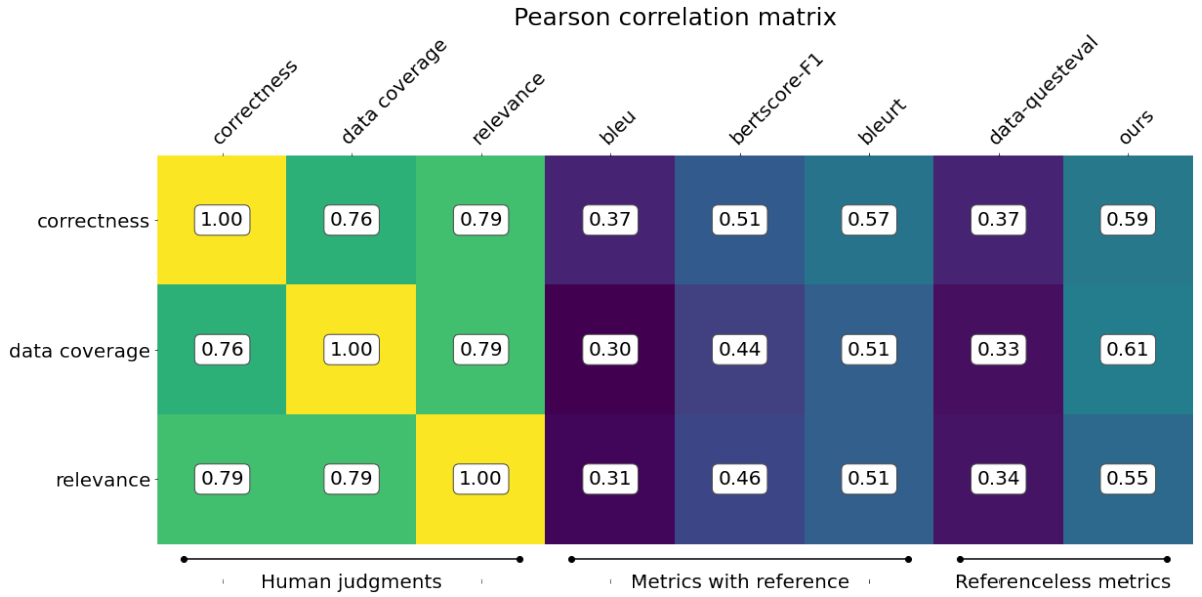
(Samuel Beckett, larger than, ***André the Giant***).

Inverted negatives are added to the mix of artificial negatives in the batches to make the model robust to inversion

Evaluation

Correlations between our metric and human scores for 2,848 generated texts (16 systems, 178 outputs) from WebNLG 2020 annotated with human judgments for:

- Data Coverage: Does the text include descriptions of all predicates present in the input?
- Relevance: Does the text describe only triples present in the graph?
- Correctness: For predicates in the graph, does the text correctly describe their arguments?
- Text Structure: Is the text grammatical, wellstructured, written in acceptable English?
- Fluency: Does the text progress naturally and form a coherent, easy-to-understand whole?



Best-performing referenceless metric

Better than BLEURT, the previous best-performing reference based metric

Semantic Evaluation of Multilingual Data-to-Text Generation via NLI Fine-Tuning: Precision, Recall and F1 scores

W. Soto-Martinez, Y. Parmentier and C. Gardent. Semantic Evaluation of Multilingual Data-to-Text Generation via NLI Fine-Tuning: Precision, Recall and F1 scores. In Submission

Goals

Multilingual

- High Resource Languages: English, Russian
- Low Resource Languages: Breton, Irish, Maltese, Welsh, Xhosa

Fine-grained evaluation of semantic similarity

- Quantifying Under-Generation (***Omissions***)
- Quantifying Over-Generation (***Additions***)

Based on Natural Language Inference (NLI)

Method based on Natural Language Inference (NLI)

Precision ($KG \models Text$)

How many of the facts expressed by the text can be inferred from the graph ?

$$\frac{\text{Nb of Correct facts Expressed by Text}}{\text{Nb of facts expressed by the text}}$$

Low Precision indicates additions

Recall ($Text \models KG$)

How many of the facts in the graph can be inferred from the text ?

$$\frac{\text{Nb of Correct facts Expressed by Text}}{\text{Nb of facts in graph}}$$

Low recall indicates omissions

Graph			
Alan Bean | birthDate | 1932-03-15 Alan Bean | almaMater | UT Austin, B.S. 1955 Alan Bean | birthPlace | Wheeler, Texas			
Texts	**Precision**	**Recall**	**Errors**
Alan Bean was born on March 15, 1932.	1/1	1/3	2O
Alan Bean was born in Wheeler, Texas and was in the Apollo 12 mission.	1/2	1/3	1A, 2O
Alan Bean was born on March 15, 1932 in Wheeler, Texas. He received a Bachelor of Science degree at the University of Texas at Austin in 1955.	3/3	3/3	None

Regression model

- Estimates the ***degree*** to which the text/graph is faithful to the graph/text
- Fine tuned on data created to capture different combinations of precision and recall
- Label: entailment weights of the classification head

Training Data

1.77M ***(KG, Text, Precision, Recall)*** quadruples across 6 languages with a balanced and diverse distribution of P and R combinations

Derived from the WebNLG dataset of (KG, English Text) pairs

We derive non aligned (g', t) pairs from $(g, t) \in WebNLG$ by pairing the text t with graphs g' which

- are sub-graphs or super graphs of g
- or where a triple contained in g is modified

We then compute ***precision*** and ***recall*** for each new (g', t) pair based on the number of added, removed or modified triples.

We machine translate the ***English text*** into the 5 target languages using the NLLB model and filtering using language identification scores and a cosine threshold (0.60) on LaBSE embeddings.

Models

mDeBERTa base multilingual NLI model fine-tuned on the training data

- **MultiFF**: Full fine-tuning of the NLI Base model on all languages together.
- **MultiLR**: LoRA on top of the NLI-Base model on all languages together.
- **MonoLR**: Lora on top of the NLI-Base model for each language individually.

Baselines

- Data-QuestEval(DQE): Question-Based
- NLI Base (NB, Dusek and Kasner 2020): NLI-Based Classification Model, English only
- FactSpotter(FS): NLI-based Classification Model, English only

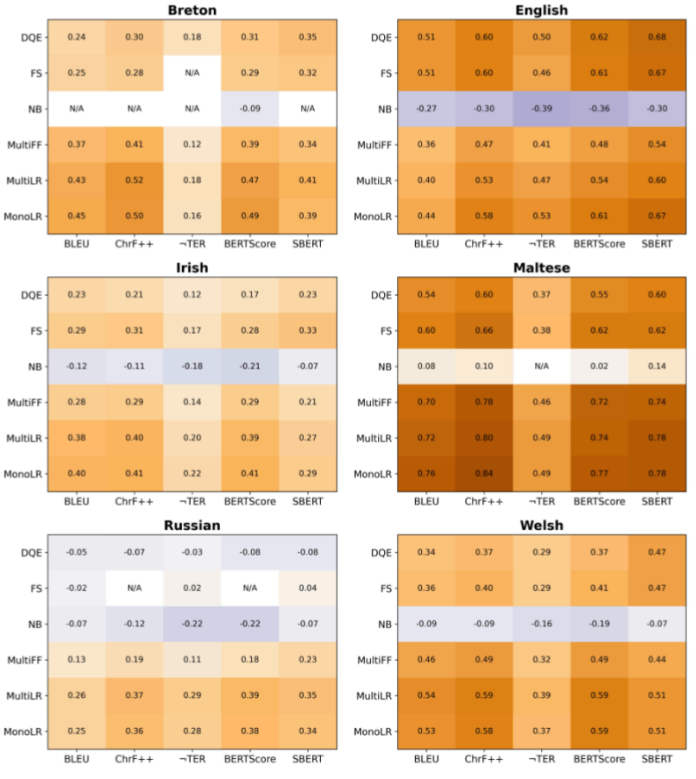
Evaluation

- Correlation with automatic metrics (7 languages)
  In the absence of reference, can our model be used as a substitute for reference-based metrics ?
- Correlation with human judgments (6 languages)
- Graph/text retrieval accuracy (7 languages).

Correlation with Automatic Metrics

Data (7L-Auto): 4,461 graphs, 148K Texts in 7 languages

- All graphs from the WebNLG testsets
- All the texts generated from these graphs by participant systems of the WebNLG 2017, 2020 and 2023 Shared Tasks
  - Grammar-based- and template-based approaches, statistical MT, neural models trained from scratched and fine-tuned pretrained models
  - Covers a wide spectrum of errors and quality level



Correlation with Automatic Metrics

Fine-tuning matters

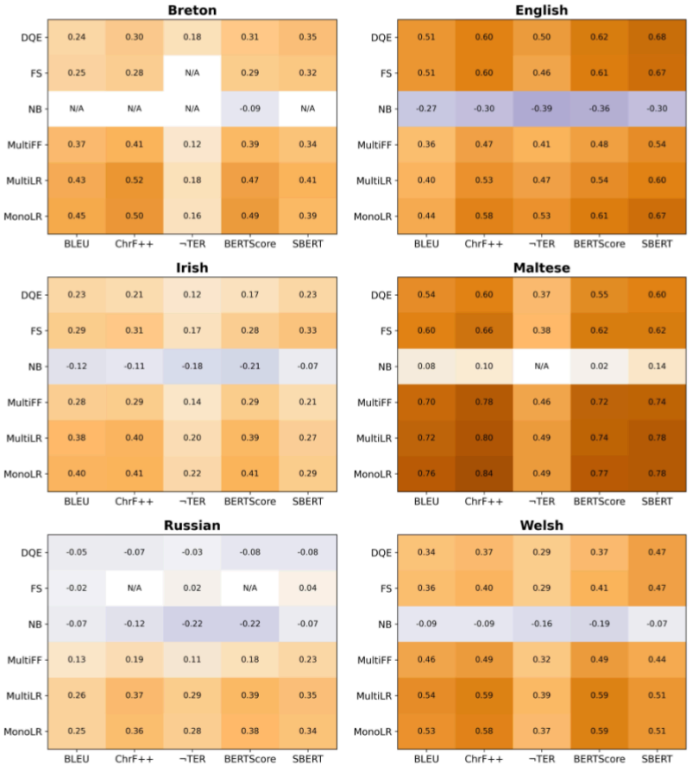
- Simply using off-the shelf models as proposed in Kasner et al. (NB model) does not suffice

Strong performance on English

- almost on par with English trained models (DQE, Factspotter)

Good results on other languages

- The monolingual Lora models outperform all three baselines on all other languages



Correlation with Human Annotations

Human judgements from WebNLG 2017, 2020 and 2023

- We reconstruct an F1 score from the human judgments provided by these datasets (product of three criteria for 2020 and Harmonic mean of binary scores for lack of addition and omission for 2023)

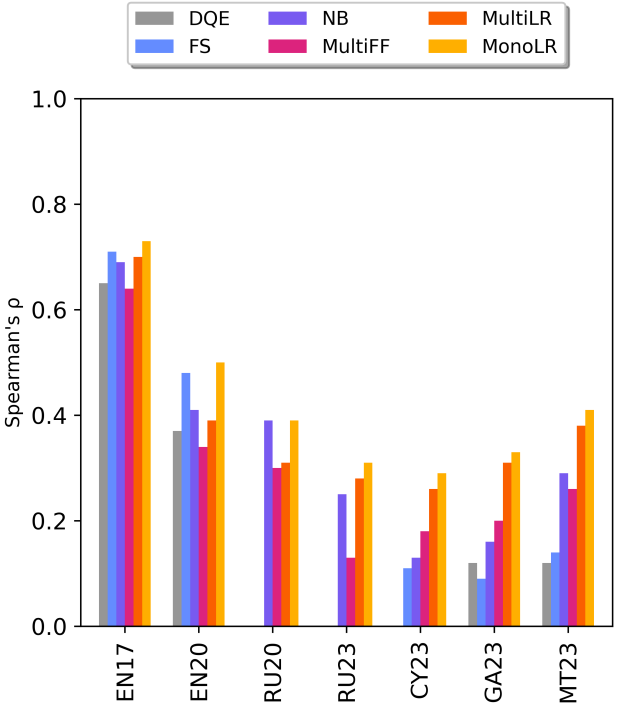
Data (4L-RP-Human): 50 graph-text pairs for 4 target languages (English, Maltese, Russian, Welsh) with a balanced distribution of precision and recall scores by our best performing model.

- The human annotators were provided with a text and a graph and asked to answer, using a scale of 1 to 5 (None, Few, Half, Most, All), the following questions:
  - Precision: *How many Triples from the text can you find in the Table?*
  - Recall: *How many Triples from the table can you find in the Text?*

Correlation with Human Annotations (WebNLG 2017, 2020, 2023)

Mixed results

- Best correlation for WebNLG 2017
- The MonoLR model outperforms the three baselines
- The gap with the English-based baselines increases for the other languages



Correlation with Human Annotations (4L-RP)

Language	Annotators	Precision		Recall		F1
		Fleiss κ	ρ	Fleiss κ	ρ	ρ
English	4	0.47	0.68	0.47	0.63	0.70
Maltese	3	0.29	0.38	0.49	0.30	0.47
Russian	2	0.32	0.63	0.39	0.52	0.67
Welsh	4	0.37	0.60	0.50	0.81	0.70

- Strong Spearman correlation for all three metrics for English, Russian and Welsh
- Moderate correlation for Maltese

  The approach adequately measures omissions (recall), addition (precision) and semantic faithfulness (F1).

MuCAL: Contrastive Alignement for Preference-Driven KG-to-Text Generation

Y. Song and C. Gardent. MuCAL: Contrastive Alignement for Preference-Driven KG-to-Text Generation. In Submission

MuCAL: Contrastive Alignment for Preference-Driven KG-to-Text Generation

- Multilingal KG-Text Encoder
  - Bi- and Cross-encoder
  - Trained on multilingual Graph/Text data
  - Using contrastive learning
- Used as a ranker to create preference data *(KG, chosen text, rejected text)*
- Train KG-to-Text model on preference data
  - compare with other KG-Text metrics

Graph/Text Training and Test Data

Dataset	Description	# KG/Text Pairs
Source Datasets		
WebNLG-Train	Gold	14,878
KELM-Q1	Silver	18,723
WebNLG-Test	Gold	1,779
KELM-Test	Gold	3,437
Training Sets		
EN-Train	KELM-Q1 + WebNLG-Train	33,601
Multi-Train-Silver	EN-Train + Translations	201,606
Test Sets		
Multi-Test-1K	1K (KELM-Test + WebNLG-Test) + Translations	6,000
Multi-WebNLG-Test	WebNLG-Test + Translations	10,674
Multi-Test-1K-Corr	Multi-Test-1K + Corrupted Graphs	10,800

Soft Nearest Neighbor Loss

Uses all positive (all 6 verbalisations of a graph) and negative points in the batch

$$-\sum_{i \in I} \log \left(\frac{\exp\left(\sum_{lg \in L} \mathrm{sim}(\mathrm{t}_i^{lg}, \mathrm{g}_i)/\tau\right)}{\sum_{j \in I} \exp \sum_{lg \in L} \left(\mathrm{sim}(\mathrm{t}_i^{lg}, \mathrm{g}_j)/\tau\right)} \right)$$

Models

Our models

Cross-encoder

- MultiMPNet fine-tuned as a cross encoder on the alignment data

Bi-encoder

- MultiMPNet fine-tuned as a bi-encoder on the alignment data

Variants

- Different batch-Size (8, 16, 32)
- Mono or bidirectional
- Without or with Hard Negatives (1, 2, 4)

Baselines

- MultiMPNet, a text based multilingual bi-encoder
- BGE-M3, the current multilingual SOTA embedding model for text
- EREDAT, a state-of-the-art KG-English text alignment

Evaluation

Retrieval on 3 test-sets of increasing complexity

1K (Easy)

- KG sampled from WebNLG and Kelm
- English text and translations into 5 target languages
- Little overlap in terms of properties and entities

WebNLG (Medium Hard)

- 1,779 graphs of the WebNLG test set for English
- WebNLG English verbalisations and translations into 5 target languages
- High overlap

1K-Corr (Hard)

- 1K graphs, 6K texts
- Each text is paired with its graph and n corrupted graphs
- The corrupted graphs are similar to the correct graph

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR
DPO Models										
**BE-MPNet-Hard2*	95.60	97.30	96.10	97.40	80.33	86.92	81.62	87.65	**73.50**	**84.55**
**CE-MPNet (bs4)*	96.40	97.51	96.60	97.53	**85.39**	**90.52**	**86.23**	**91.20**	24.10	55.30
Baselines										
BE-MultiMPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
BE-BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
BE-EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
Ablation Studies										
BE-MPNet (bs8)	95.70	97.53	96.10	97.79	79.60	86.61	81.06	88.04	41.90	65.66
BE-MPNet (bs16)	96.60	98.14	97.60	98.69	82.18	88.37	83.08	89.50	43.40	67.38
BE-MPNet (bs32)	96.10	97.66	97.60	98.69	83.53	89.34	84.94	90.68	46.40	69.53
BE-MPNet-BiDir	**96.90**	**98.34**	**98.10**	**98.98**	84.20	89.68	85.33	90.90	49.60	71.44
BE-MPNet-Hard1	95.00	96.99	96.70	97.90	79.26	86.33	81.84	88.05	69.90	82.75
BE-MPNet-Hard4	94.90	96.85	94.20	96.11	78.70	85.61	78.81	85.77	69.60	81.76

Table 2: Model Performance Comparison on Test Sets for monolingual tasks (English). * Our final model selections for preference learning. BE: Bi-Encoder, CE: Cross-Encoder, G2T: Graph-to-Text Retrieval, T2G: Text-to-Graph Retrieval, Accuracy (Acc), Mean Reciprocal Rank (MRR). The batch size (bs) for all BE models is 32 unless it is explicitly stated.

Improvement over baselines

Text based multilingual encoders under-perform on the hard test sets (WebNLG, 1K-Corr)

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR
DPO Models										
**BE-MPNet-Hard2*	95.60	97.30	96.10	97.40	80.33	86.92	81.62	87.65	**73.50**	**84.55**
**CE-MPNet (bs4)*	96.40	97.51	96.60	97.53	**85.39**	**90.52**	**86.23**	**91.20**	24.10	55.30
Baselines										
BE-MultiMPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
BE-BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
BE-EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
Ablation Studies										
BE-MPNet (bs8)	95.70	97.53	96.10	97.79	79.60	86.61	81.06	88.04	41.90	65.66
BE-MPNet (bs16)	96.60	98.14	97.60	98.69	82.18	88.37	83.08	89.50	43.40	67.38
BE-MPNet (bs32)	96.10	97.66	97.60	98.69	83.53	89.34	84.94	90.68	46.40	69.53
BE-MPNet-BiDir	**96.90**	**98.34**	**98.10**	**98.98**	84.20	89.68	85.33	90.90	49.60	71.44
BE-MPNet-Hard1	95.00	96.99	96.70	97.90	79.26	86.33	81.84	88.05	69.90	82.75
BE-MPNet-Hard4	94.90	96.85	94.20	96.11	78.70	85.61	78.81	85.77	69.60	81.76

Table 2: Model Performance Comparison on Test Sets for monolingual tasks (English). * Our final model selections for preference learning. BE: Bi-Encoder, CE: Cross-Encoder, G2T: Graph-to-Text Retrieval, T2G: Text-to-Graph Retrieval, Accuracy (Acc), Mean Reciprocal Rank (MRR). The batch size (bs) for all BE models is 32 unless it is explicitly stated.

Degradation on harder test sets: - 1K > WebNLG > 1K-Corr

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR
DPO Models										
**BE-MPNet-Hard2*	95.60	97.30	96.10	97.40	80.33	86.92	81.62	87.65	**73.50**	**84.55**
**CE-MPNet (bs4)*	96.40	97.51	96.60	97.53	**85.39**	**90.52**	**86.23**	**91.20**	24.10	55.30
Baselines										
BE-MultiMPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
BE-BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
BE-EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
Ablation Studies										
BE-MPNet (bs8)	95.70	97.53	96.10	97.79	79.60	86.61	81.06	88.04	41.90	65.66
BE-MPNet (bs16)	96.60	98.14	97.60	98.69	82.18	88.37	83.08	89.50	43.40	67.38
BE-MPNet (bs32)	96.10	97.66	97.60	98.69	83.53	89.34	84.94	90.68	46.40	69.53
BE-MPNet-BiDir	**96.90**	**98.34**	**98.10**	**98.98**	84.20	89.68	85.33	90.90	49.60	71.44
BE-MPNet-Hard1	95.00	96.99	96.70	97.90	79.26	86.33	81.84	88.05	69.90	82.75
BE-MPNet-Hard4	94.90	96.85	94.20	96.11	78.70	85.61	78.81	85.77	69.60	81.76

Table 2: Model Performance Comparison on Test Sets for monolingual tasks (English). * Our final model selections for preference learning. BE: Bi-Encoder, CE: Cross-Encoder, G2T: Graph-to-Text Retrieval, T2G: Text-to-Graph Retrieval, Accuracy (Acc), Mean Reciprocal Rank (MRR). The batch size (bs) for all BE models is 32 unless it is explicitly stated.

Hard negatives help on hard test sets

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR	Acc	MRR
DPO Models										
**BE-MPNet-Hard2*	95.60	97.30	96.10	97.40	80.33	86.92	81.62	87.65	**73.50**	**84.55**
**CE-MPNet (bs4)*	96.40	97.51	96.60	97.53	**85.39**	**90.52**	**86.23**	**91.20**	24.10	55.30
Baselines										
BE-MultiMPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
BE-BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
BE-EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
Ablation Studies										
BE-MPNet (bs8)	95.70	97.53	96.10	97.79	79.60	86.61	81.06	88.04	41.90	65.66
BE-MPNet (bs16)	96.60	98.14	97.60	98.69	82.18	88.37	83.08	89.50	43.40	67.38
BE-MPNet (bs32)	96.10	97.66	97.60	98.69	83.53	89.34	84.94	90.68	46.40	69.53
BE-MPNet-BiDir	**96.90**	**98.34**	**98.10**	**98.98**	84.20	89.68	85.33	90.90	49.60	71.44
BE-MPNet-Hard1	95.00	96.99	96.70	97.90	79.26	86.33	81.84	88.05	69.90	82.75
BE-MPNet-Hard4	94.90	96.85	94.20	96.11	78.70	85.61	78.81	85.77	69.60	81.76

Table 2: Model Performance Comparison on Test Sets for monolingual tasks (English). * Our final model selections for preference learning. BE: Bi-Encoder, CE: Cross-Encoder, G2T: Graph-to-Text Retrieval, T2G: Text-to-Graph Retrieval, Accuracy (Acc), Mean Reciprocal Rank (MRR). The batch size (bs) for all BE models is 32 unless it is explicitly stated.

Larger batch size helps

Direct Preference Optimisation for KG-to-Text Generation

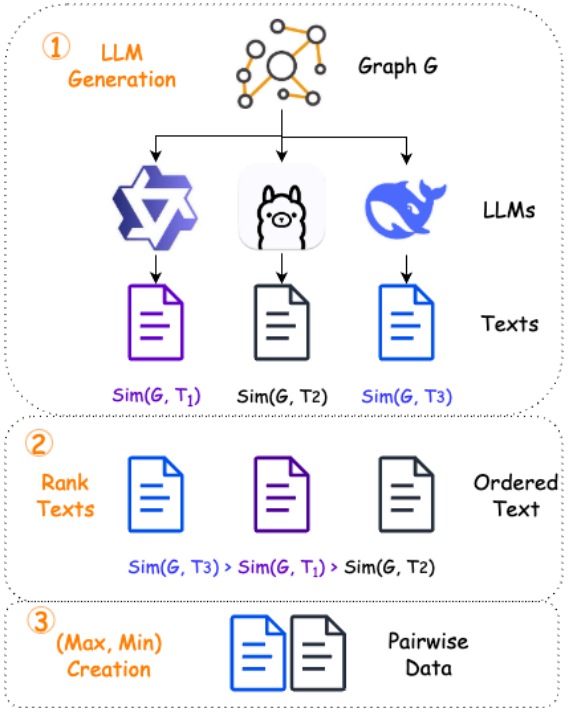
1. Create preference data
(KG, good output, bad output)

2. Fine tune KG-to-Text model on KG/Text data

3. DPO optimisation on preference data

Creating Preference Data

- Use LLMs to verbalise the graph
- ***Compute the similarity between the graph and each generated text*** (4 KG/Text scoring metrics)
- Rank the texts accordingly
- Select the texts with the highest and lowest similarity scores to create preference pairs.



Creating Preference Data

Generating Candidate Texts

Graphs from Kelm-Q1

LLMs: Qwen2.5 7B/14B/32B Instruction Variants, DeepSeek-v3, r1-distill-Qwen-7B, Llama-3-8-Instruct

- Three shots from KELM test set
- 6 texts/graph

Scoring and Ranking Candidates

3 KG/Text similarity metrics

- EREDat
- FactSpotter
- Data Quest-Eval

Creating Preference Triples

We maximise the scoring gap between preferred and dispreferred text

(graph, top-ranked text, bottom-ranked text)

DPO Training

Step 1: Fine tune Qwen2.5-1.5B Instruct on Kelm-Q1

Step 2: Optimise on preference data using DPO objective

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(t_{\text{C}},t_{\text{R}})\sim\mathcal{D}_{\text{pref}}} \log \sigma \left(\beta \, \Delta_\theta(G, t_C, t_R)\right)$$

$$\Delta_\theta(G, t_C, t_R) = \log \frac{\pi_\theta(t_C|G)}{\pi_{\text{ref}}(t_C|G)} - \log \frac{\pi_\theta(t_R|G)}{\pi_{\text{ref}}(t_R|G)}$$

(π_{ref}) is the instruction-tuned reference policy (our fine-tuned model)
(π_θ) is the training policy (the model we want to learn)
$(\beta = 0.1)$ controls the KL regularization strength
(σ) is the sigmoid function.
t_C, chosen
t_R, rejected

Models

5 DPO models

- 2 trained on preference data created using our 2 KG-Text alignment models (Bi- and Cross-encoder)
- 3 trained on preference data created using MuCAL, EREdat, FactSpotter and DataQuestEval

2 LLMs

- Zero- and 3-shot Qwen

Qwen fine tuned on Kelm-Q1

Evaluation

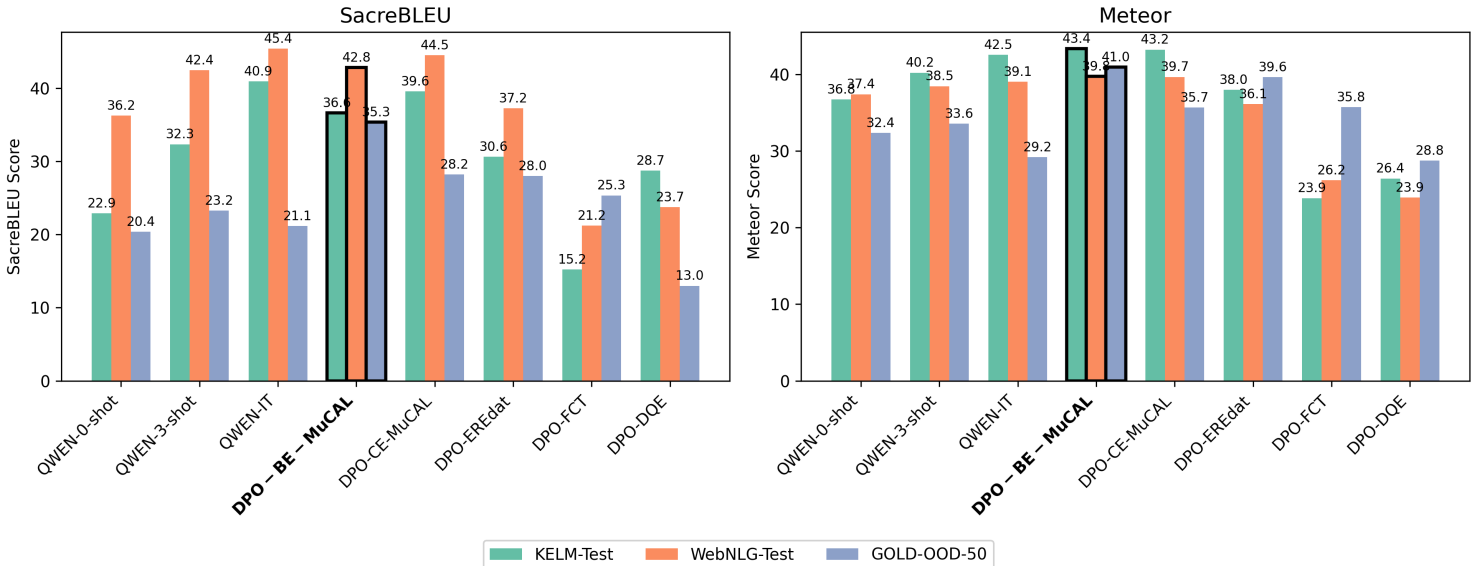
Test Sets

- KELM-Test: In-domain
- WebNLG: Public
- GOLD-OOD-50: Out-of-Domain

Metrics:

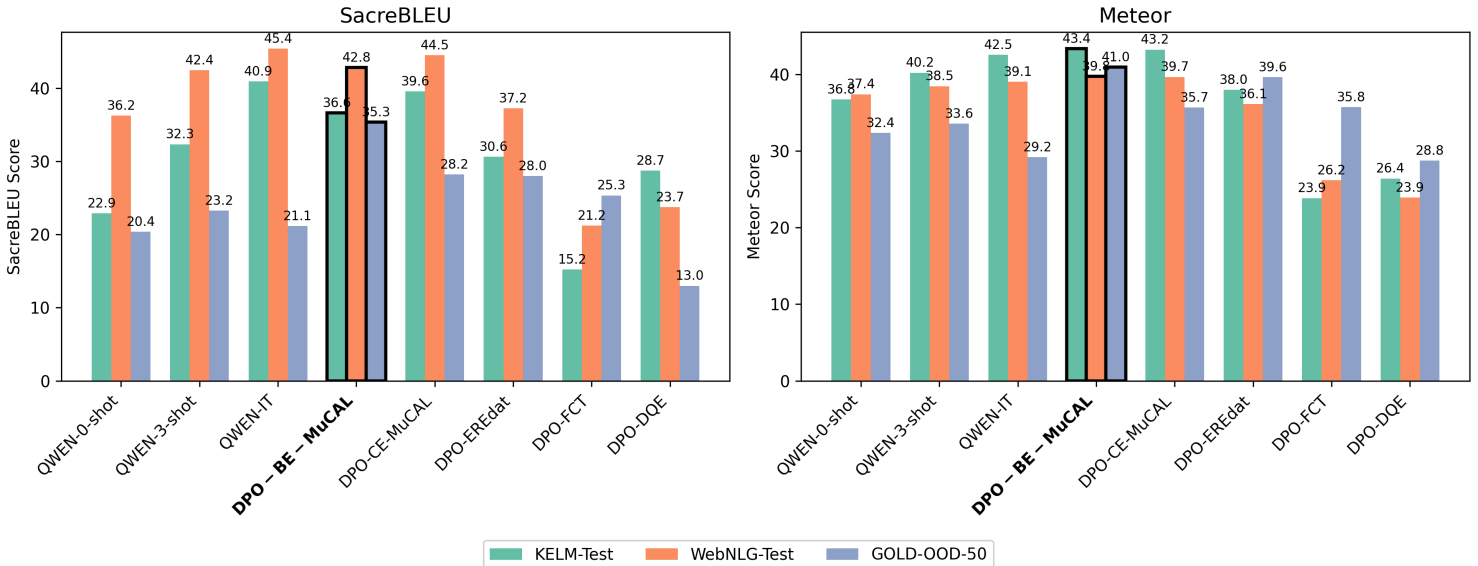
- Reference-less metrics: EREdat, FactSpotter, Data Quest-Eval
- Reference-based metrics: SacreBLEU, METEOR, TER, ...

Results



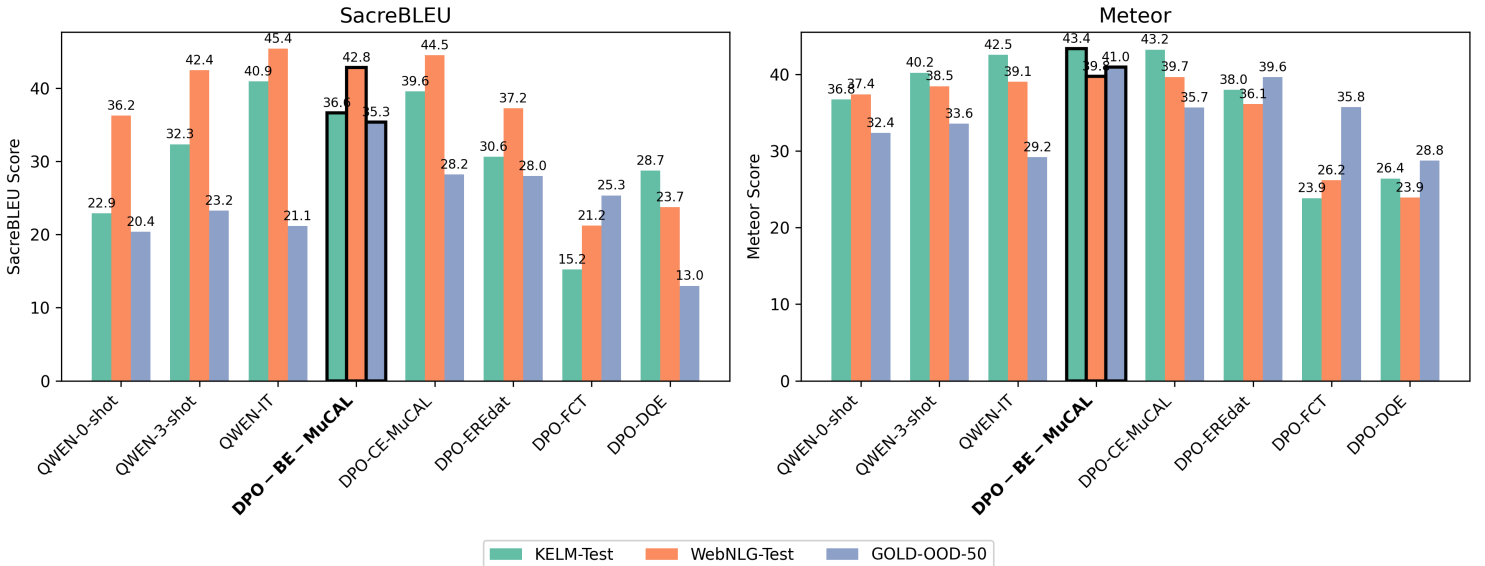
DPO models generalise better to OOD data (GOLD-OOD-50)

Results



DPO models generalise better: they paraphrase the reference (higher METEOR scores)

Results



The bi-encoder outperforms the cross-encoder on OOD data (Hard negatives are important)

Better Factual Consistency

Model	BLEU	Meteor	ChrF	TER	BertScore	Bleurt	Eredat	Facts	Parent	Quest-Eval	Sescore2
Prompting Baselines											
QWEN-0-shot	22.89	36.75	16.84	91.40	93.64	75.22	84.69	67.27	34.25	69.10	-6.33
QWEN-3-shot	32.33	40.23	65.89	54.49	95.04	78.95	88.72	82.22	45.09	71.85	-3.24
Instrution Tuning											
QWEN-IT	**40.91**	42.55	69.77	**42.99**	**95.77**	**81.67**	90.40	56.93	50.72	71.64	**-1.69**
DPO Variants											
DPO-Mpnet-hard2	36.60	**43.37**	69.83	54.35	95.14	78.27	92.38	91.70	52.18	71.65	-2.70
DPO-Eredat	30.62	37.99	59.76	100.06	93.92	76.20	**92.59**	91.89	49.51	71.72	-2.70
DPO-FactSpotter	15.23	23.85	11.30	772.00	90.10	64.45	80.06	**96.71**	36.01	68.69	-12.90
DPO-DQE	28.71	26.41	37.60	351.48	92.65	78.52	88.57	94.05	55.42	**74.58**	-7.16
DPO-Mpnet-CE	39.56	43.24	**69.87**	46.00	95.53	79.26	91.22	90.09	**53.03**	72.19	-2.10

DPO generates texts with higher input (graph) consistency

Conclusion

Conclusion

- Metrics which separately capture omissions and additions are useful for a finer-grained reference-less evaluation of KG-to-Text generation models
- Joint encoders are more useful for retrieval and ranking
- Future work: multilingual KG-to-Text generation and preference learning

The End