Neural Embeddings for Text and Knowledge Graphs

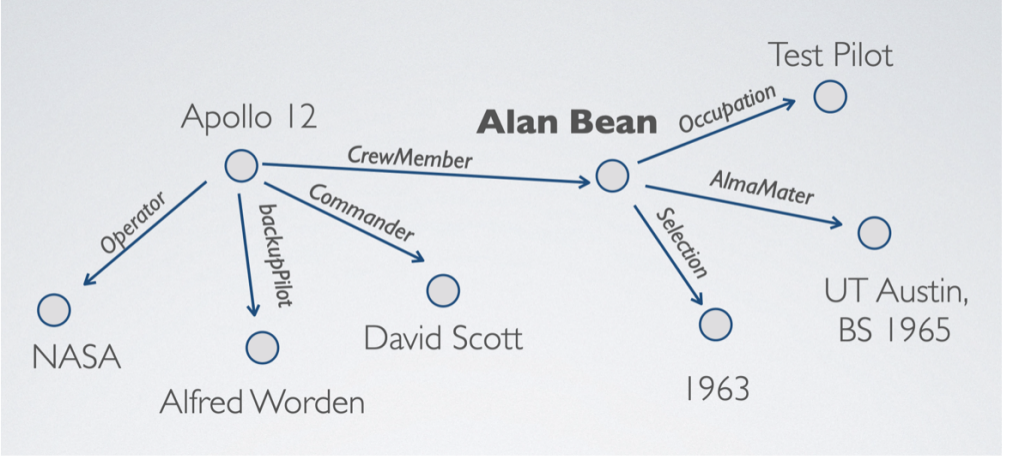
Retrieval, Evaluation, Text Generation

Claire Gardent

CNRS / LORIA, Nancy

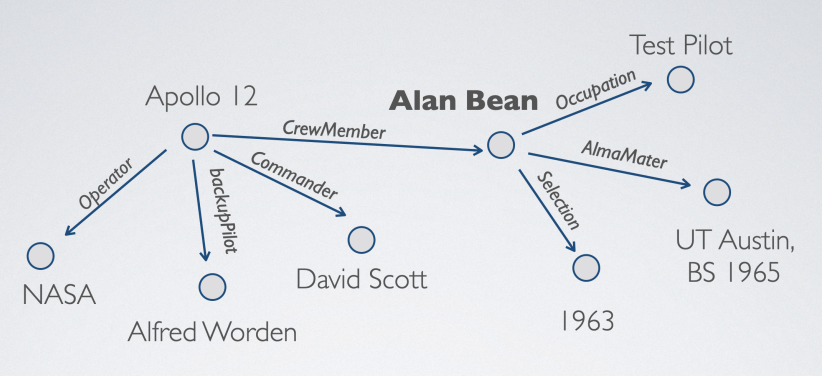
KG/Text Embeddings

Goal: Create embeddings for KGs and for text such that whenever a graph and a text convey the same or similar content, their embedding are close in the semantic space



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

Multilingual KG/Text Embeddings



[FR] Alan Bean a obtenu une licence en sciences à l'université du Texas à Austin en 1955. Il a été embauché par la NASA en 1963 et a occupé le poste de pilote d'essai. Le pilote de réserve d'Apollo 12 était Alfred Worden et le commandant était David Scott.

[ES] Alan Bean se graduó en la Universidad de Texas en Austin en 1955 con una licenciatura en Ciencias. Fue contratado por la NASA en 1963 y trabajó como piloto de pruebas. El piloto suplente del Apolo 12 era Alfred Worden y el comandante era David Scott.

[RU] Алан Бин окончил Техасский университет в Остине в 1955 году со степенью бакалавра наук. В 1963 году он был принят на работу в НАСА и работал испытательным пилотом. Запасным пилотом «Аполло-12» был Альфред Уорден, а командиром — Дэвид Скотт.

[ZH] 艾伦·宾于1955年毕业于德克萨斯大学奥斯汀分校，获理学学士学位。他于1963年受聘于美国国家航空航天局，担任试飞员。阿波罗12号的备份飞行员是阿尔弗雷德·沃登，指令长为大卫·斯科特。

Two Methods for Learning KG/Text Embeddings

Encoding

- cross- and bi-encoders

Fine-Tuning a Natural Language Inference Model

- so that matching graph/text pairs entail each other

Evaluating and Using KG/Text Embeddings

Retrieval

- Given a retrieval base of KGs (Texts), find the KG (text) that is most similar to some input text (graph).

Evaluating KG-to-Text Generation Models

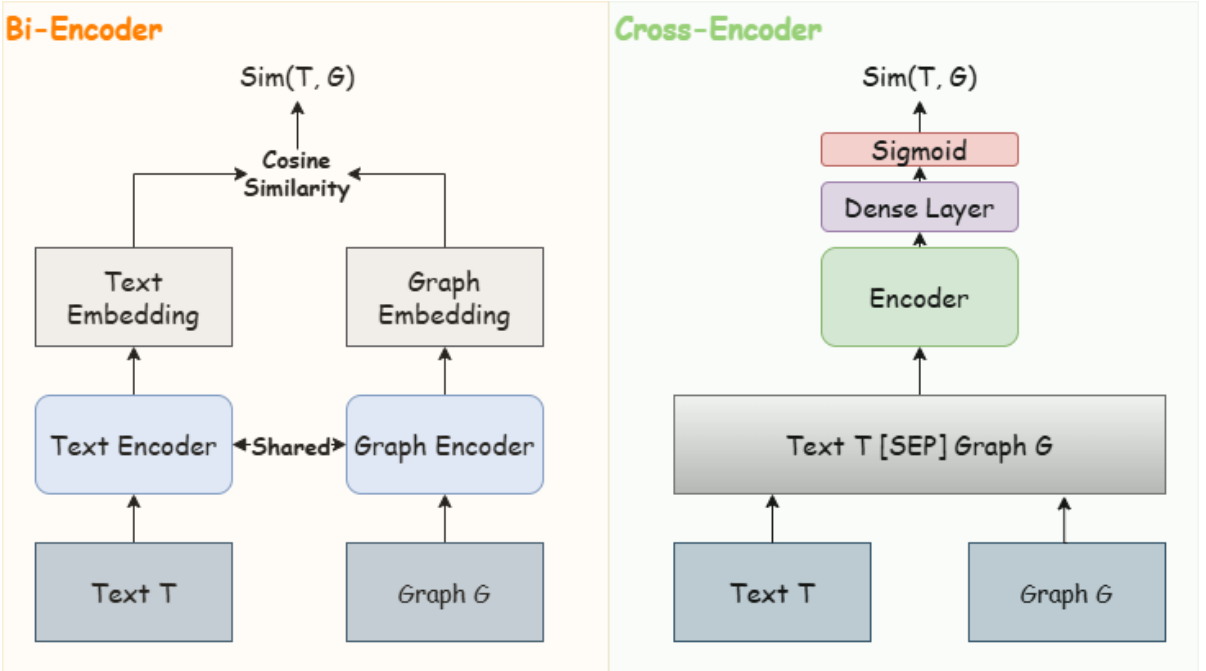
- Does the generated text convey all and only the information represented by the input knowledge graph?*

Improving KG-to-Text Generation Models

- Can we use a KG/Text similarity metrics to guide generation ?

Training KG/Text Encoders

Cross- and Bi-Encoders



The models are trained on negative and positive KG/Text pairs to maximise the similarity score of matching KG/Text pairs using a contrastive loss

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

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

Silver Data for training

Challenge: Lack of parallel data

	# (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

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.

Test Data for Retrieval

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.

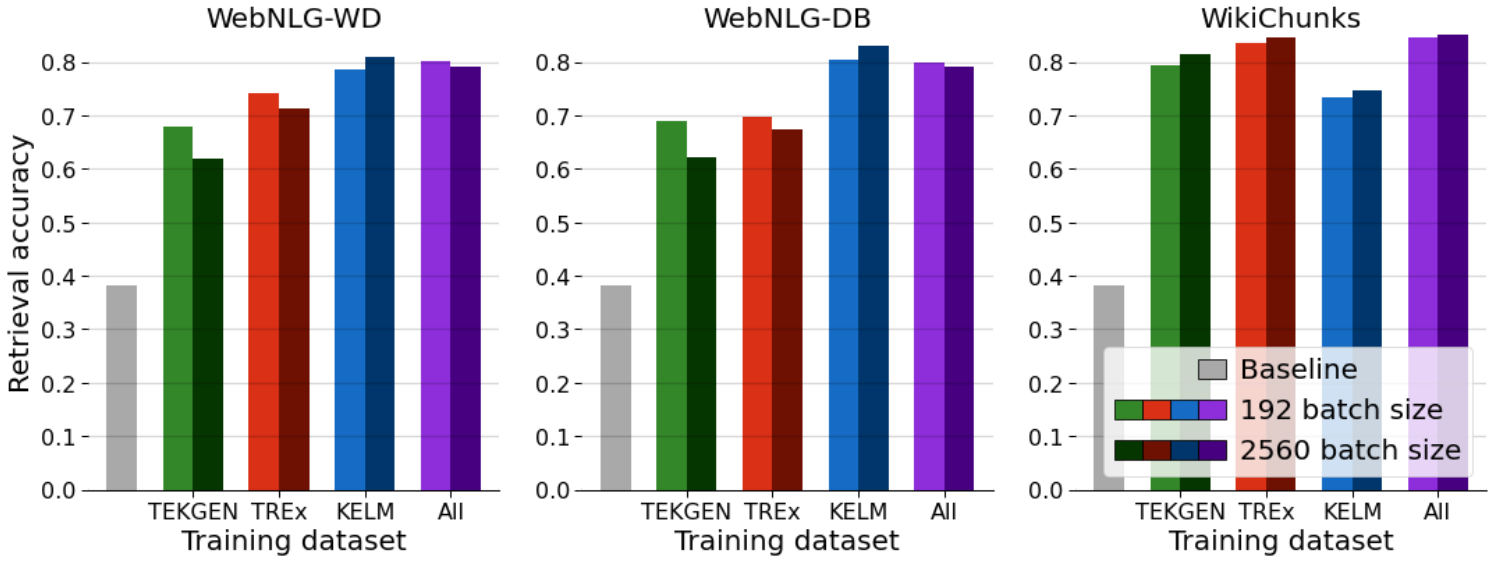
Baseline - Text Embeddings

all-mpnet-base-v2

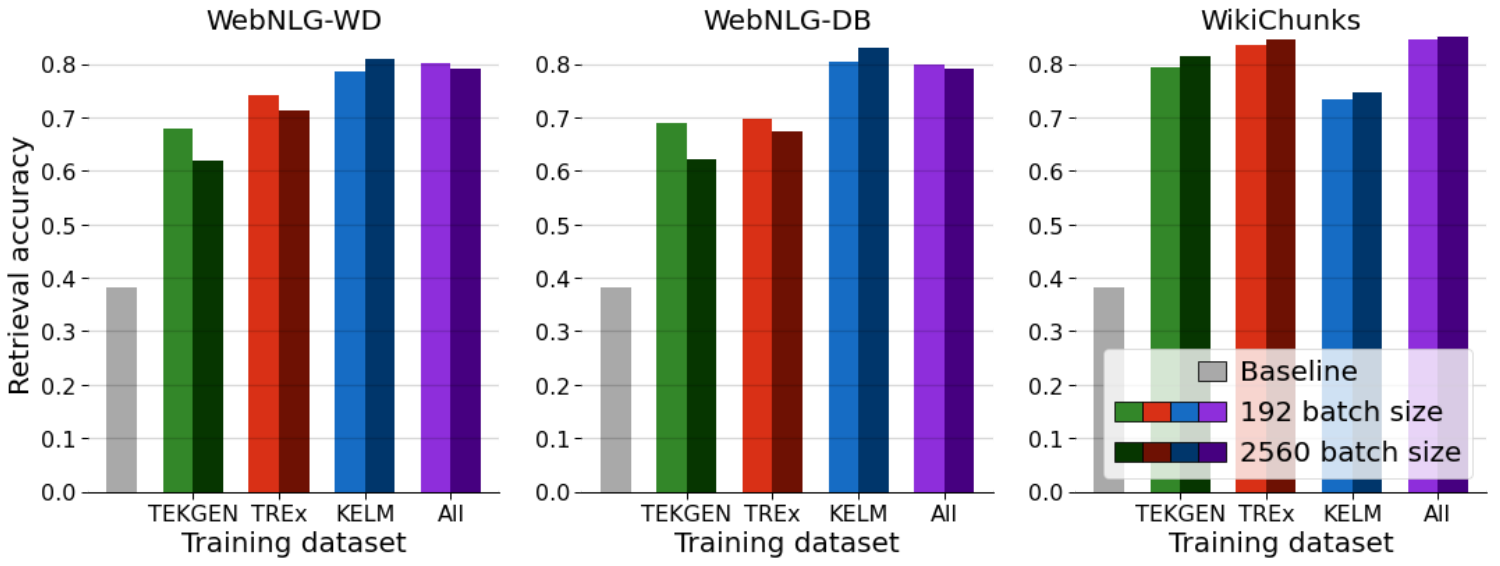
- A state-of-the-art sentence embedding model
- optimised to assess semantic similarity between texts
- used to initialise our bi-encoder

Retrieval Accuracy

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

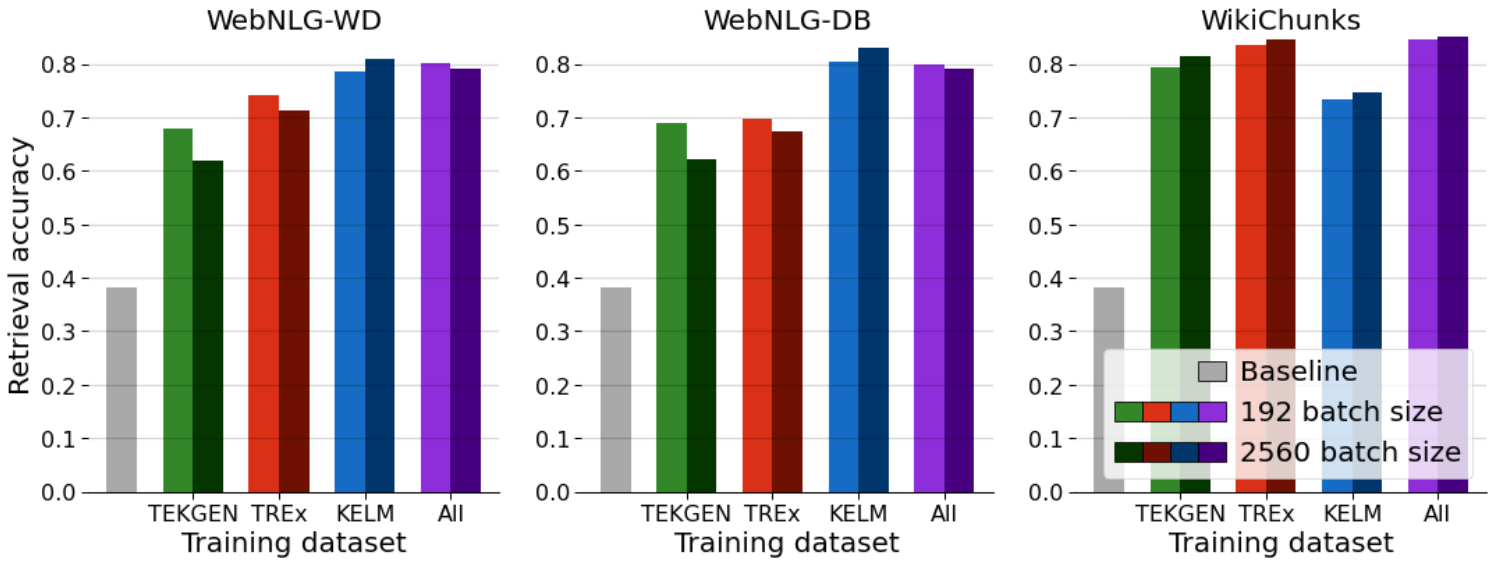


Retrieval Accuracy



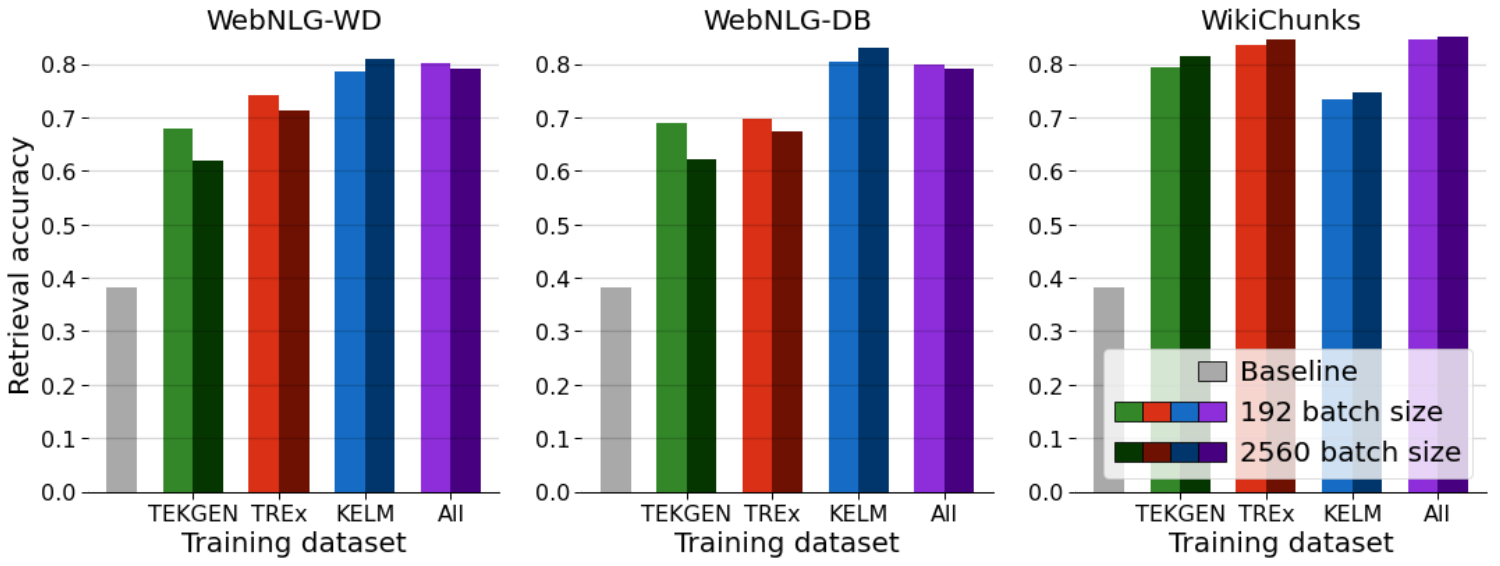
Large improvement over the baseline

Retrieval Accuracy



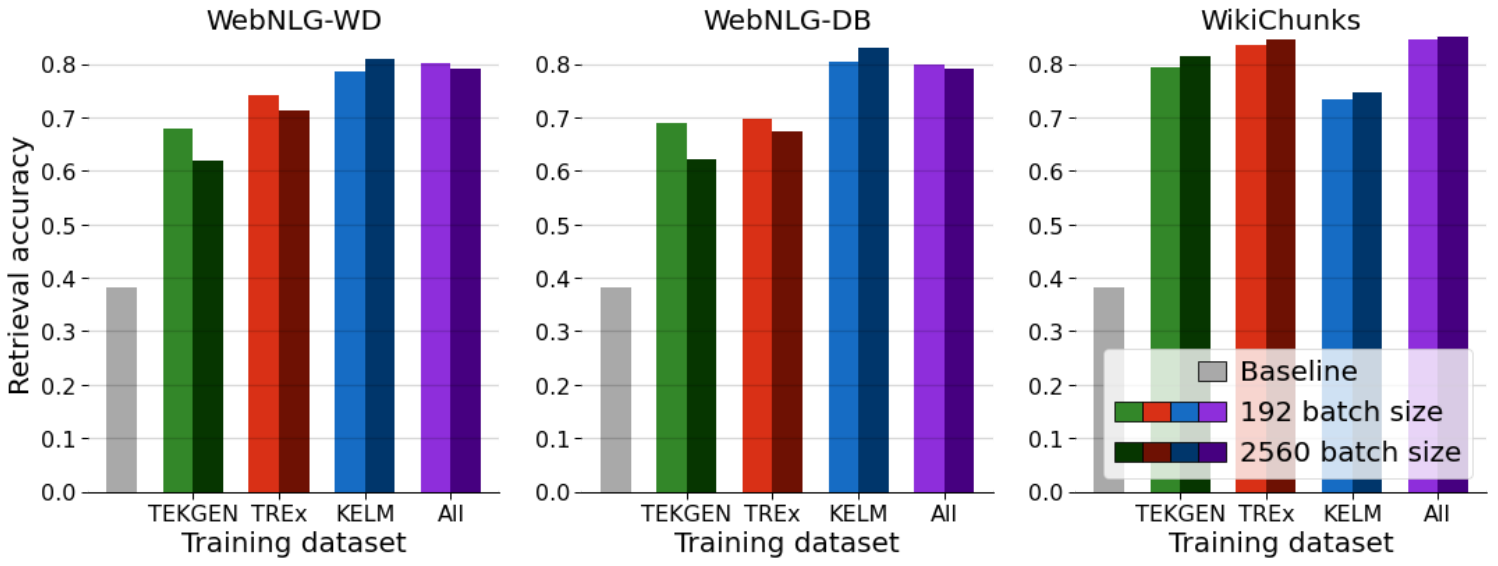
Better results when training data is similar (more or less noisy) to test data

Retrieval Accuracy



For the aligned test sets, better aligned data results in better retrieval accuracy

Retrieval Accuracy



The model generalises well to different KBs

Evaluating KG-to-Text Models

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?*

Ensemble Model

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

Inverted Negatives

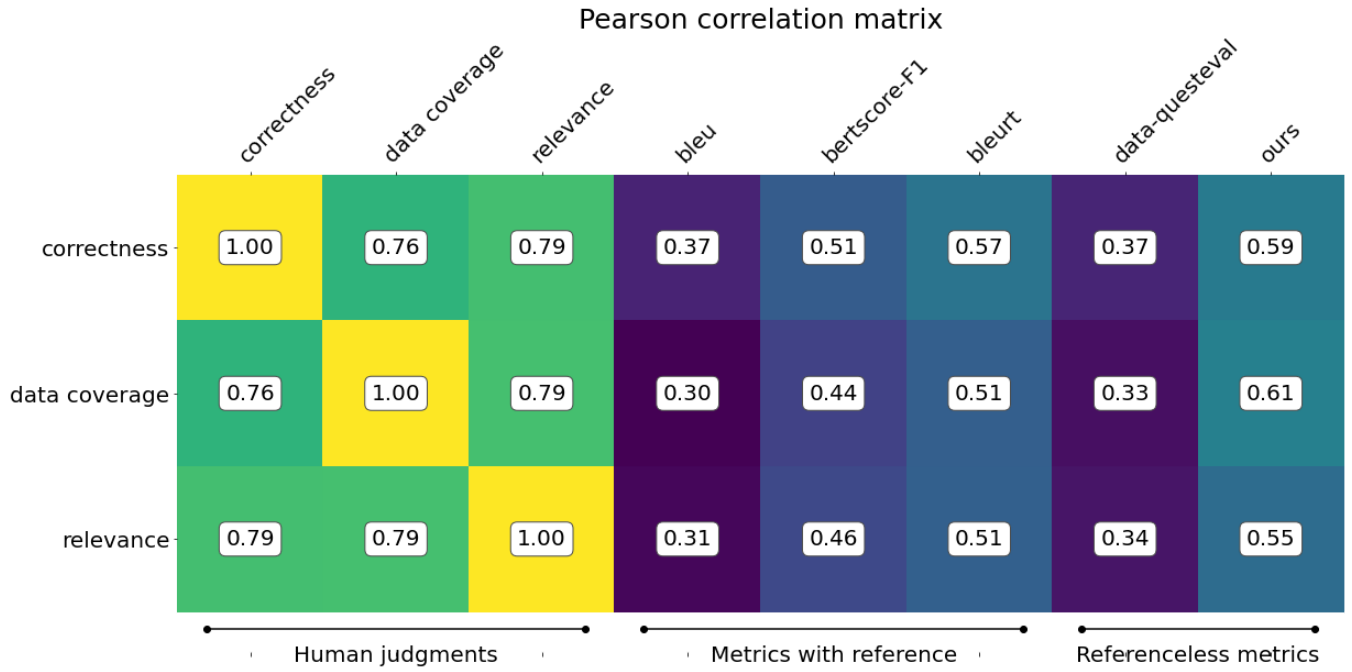
- Inverted negatives are added to the mix of artificial negatives in the batches to make the model robust to inversion

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

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?



Best-performing referenceless metric

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

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. NAACL 2025

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

- ***Multilingal*** KG-Text Encoder
- Used as a ranker to ***create preference data***
  (KG, good text, bad text)
- Use preference data to ***improve KG-to-Text model*** via preference learning

Multilingual Training Data

Each graph is associated with verbalisations in 5 languages (English, German, French, Chinese, Spanish)

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

Test Data (Retrieval)

3 test-sets of increasing complexity

1K (Easy)

- Little overlap in terms of properties and entities

WebNLG (Medium Hard)

- High overlap

1K-Corr (Hard)

- Each text is paired with its graph and n corrupted graphs
- The corrupted graphs are similar to the correct graph

Contrastive Loss Training

Positive examples: all 6 verbalisations of a graph

Negative examples: Mismatched KG/Text pairs using other graphs from the batch

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

We compare the Model with 4 approaches

Two text-based Similarity Metrics

- MultiMPNet, a text based ***multilingual*** bi-encoder
- BGE-M3, the current ***multilingual SOTA*** embedding model for text

Two KG/Text Similarity Metric

- EREDAT, a KG-English text cross encoder
- FactSpotter, a state-of-the-art KG-English text similarity metric

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR
Models Selected for Preference Learning										
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										
MPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
CLS-MPNet	91.10	94.00	91.60	94.32	65.99	76.61	62.06	74.97	29.10	57.05
BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
FactSpotter	71.10	80.74	67.70	80.52	38.90	55.46	37.27	56.90	32.70	55.77
Batch Size Variants										
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
Hard Negative Variants										
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

Text based multilingual encoders under-perform on the hard test sets

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR
Models Selected for Preference Learning										
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										
MPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
CLS-MPNet	91.10	94.00	91.60	94.32	65.99	76.61	62.06	74.97	29.10	57.05
BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
FactSpotter	71.10	80.74	67.70	80.52	38.90	55.46	37.27	56.90	32.70	55.77
Batch Size Variants										
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
Hard Negative Variants										
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

Degradation on harder test sets

$$\textit{1K} > \textit{WebNLG} > \textit{1K-Corr}$$

Results

Model Variants	Multi-Test-1K				Multi-WebNLG-Test				Multi-Test-1K-Corr	
	G2T		T2G		G2T		T2G		T2G	
	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR
Models Selected for Preference Learning										
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										
MPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
CLS-MPNet	91.10	94.00	91.60	94.32	65.99	76.61	62.06	74.97	29.10	57.05
BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
FactSpotter	71.10	80.74	67.70	80.52	38.90	55.46	37.27	56.90	32.70	55.77
Batch Size Variants										
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
Hard Negative Variants										
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

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	
	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR	R@1	MRR
Models Selected for Preference Learning										
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										
MPNet	83.20	88.98	83.20	89.16	43.28	57.17	39.91	54.67	25.00	50.25
CLS-MPNet	91.10	94.00	91.60	94.32	65.99	76.61	62.06	74.97	29.10	57.05
BGE-M3	92.90	96.09	96.00	97.77	70.49	80.55	80.04	87.69	45.90	68.53
EREDAT	95.20	97.10	96.50	98.01	76.67	84.65	82.91	89.46	41.00	66.54
FactSpotter	71.10	80.74	67.70	80.52	38.90	55.46	37.27	56.90	32.70	55.77
Batch Size Variants										
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
Hard Negative Variants										
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

Larger batch size helps

Preference-Driven KG-to-Text Generation

Y. Song and C. Gardent. MuCAL: Contrastive Alignement for Preference-Driven KG-to-Text Generation. EMNLP 2025

Direct Preference Optimisation for KG-to-Text Generation

1. Create preference data

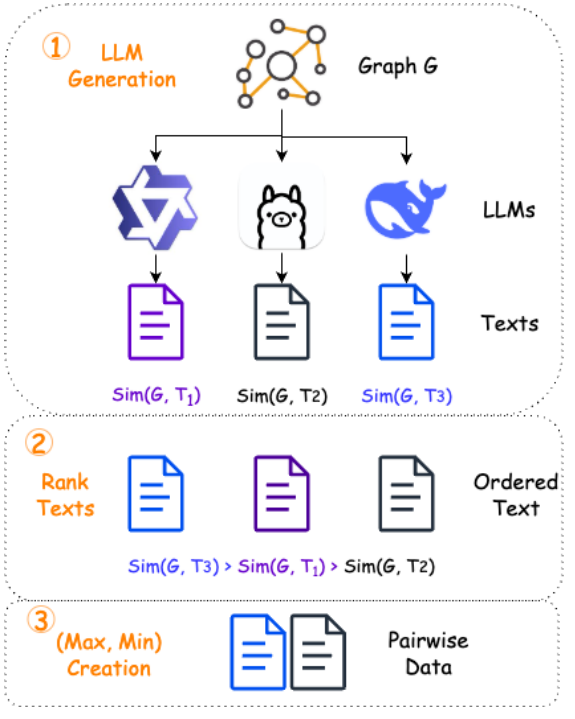
(KG, good text, bad text)

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
- ***Use MuCal and other metrics to compute the similarity between the graph and each generated text*** (4 KG/Text scoring metrics)
- Rank the texts using those metrics
- Select the texts with the highest and lowest similarity scores to create preference pairs.



Creating Preference Data

Generating Candidate Texts

6 texts/graph

- Generated by 6 LLMs: Qwen2.5 7B/14B/32B Instruction Variants, DeepSeek-v3, r1-distill-Qwen-7B, Llama-3-8-Instruct (Three shots from KELM test set)
- Graphs from Kelm-Q1

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
- Fine tuned on Kelm-Q1

Evaluation

Test Sets

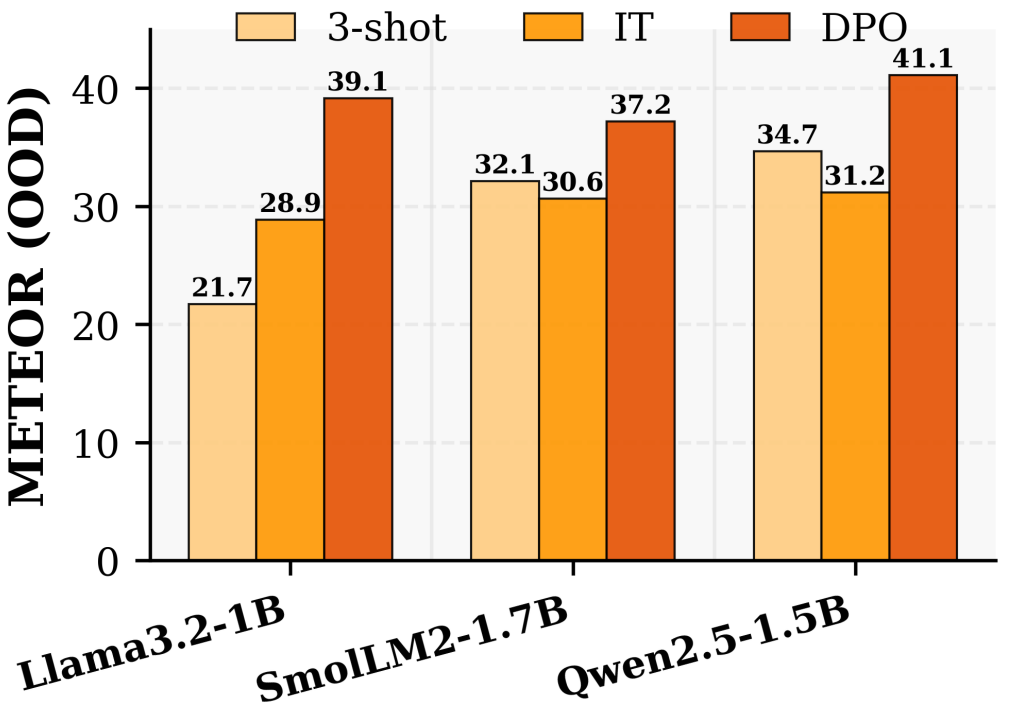
- **In-domain:** KELM-Test
- **Out-of-Domain but Public:** WebNLG
- **Out-of-Domain:** GOLD-OOD-50

Metrics:

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

DPO models generalise better to OOD data

DPO outperforms 3-Shot Prompting and Instruction Tuning



Better Factual Consistency

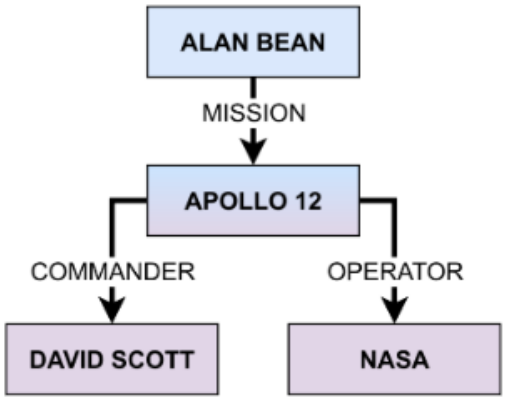
Model	BLEU	Meteor	ChrF	TER	BertScore	Bleurt	Eredat	Facts	Parent	Quest-Eval	Sescore2
Prompting Baselines											
QWEN-0-shot	22.89	36.75	16.84	91.40	93.64	75.22	84.69	67.27	34.25	69.10	-6.33
QWEN-3-shot	32.33	40.23	65.89	54.49	95.04	78.95	88.72	82.22	45.09	71.85	-3.24
Instrution Tuning											
QWEN-IT	**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

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. INLG 2024.

Quantifying Additions and Ommissions



Correct
Alan Bean was a member of Apollo 12 operated by NASA under commander David Scott

One triple missing (NASA)
Alan Bean was a member of Apollo 12 operated by NASA under commander David Scott

One triple added (birthdate)
Alan Bean, born on March 15th, was a member of Apollo 12 operated by NASA under commander David Scott

One triple added (birthdate), two triples missing (NASA, Scott)
Alan Bean, born on March 15th, was a member of Apollo 12.

Method based on Natural Language Inference (NLI)

Precision ($KG \models_{NLI} Text$)
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 ($Text \models_{NLI} KG$)
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
- Label: *entailment weights* of the classification head
- Fine tuned on data created to capture different combinations of precision and recall

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

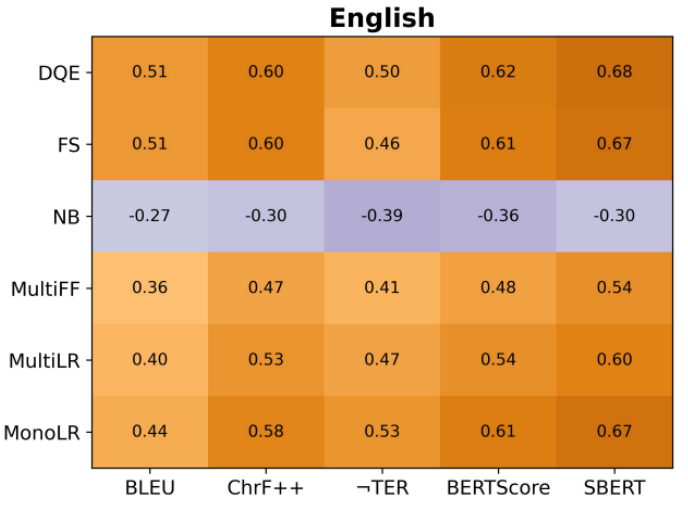
Evaluation

- 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

- 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

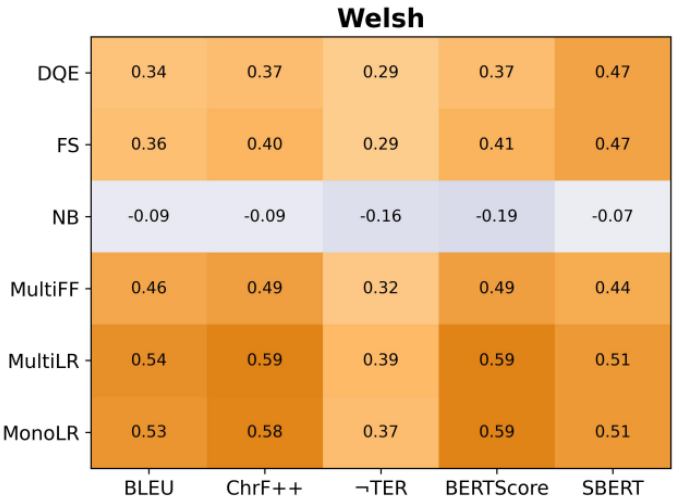


Fine-tuning matters - Simply using off-the shelf models as proposed in Kasner et al. (NB model) does not suffice

Correlation with Automatic Metrics

The trained models (3 lower rows) have better correlation with automatic metrics than the three baselines (top rows).

The LoRA models (MultiLR, MonoLR) are best.



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)

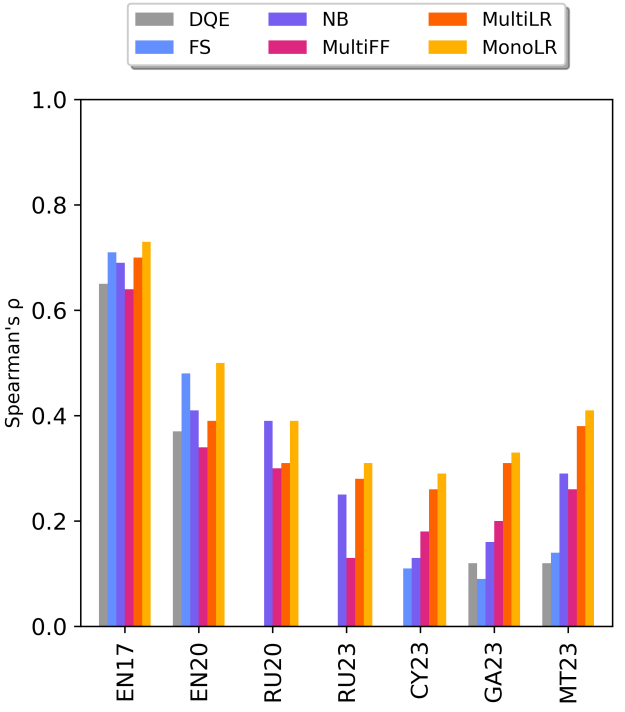
Data (4L-RP-Human): 50 graph-text pairs for 4 target languages (English, Maltese, Russian, Welsh) with a balanced distribution of precision and recall scores by our best performing model.

- 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

Takeaways

- *Reference-less*, *multilingual metric* for the evaluation of KG-to-Text models
- *Fine-grained Metrics* which evaluates omissions (low precison), additions (low recall) and semantic faithfulness (high F1).
- *Correlates well* with automatic metric and human judgements

Thanks!

Anja Belz, Thiago Castro-Ferreira, Liam Cripwell, Albert Gatt, Nikolai Ilinyskh, Anna Nikiforovskaja, Chris van der Lee, Simon Mille, Diego Moussalem, Yannick Parmentier, Laura Perez-Beltrachini, Anastasia Shimorina

Yifei Song

William Soto-Martinez

Funding

ENACT
Make Grand Est a European leader in Artificial Intelligence

The End