Verbalising Knowledge Graphs into Multiple Languages

Models, Evaluation and Generalisation

Claire Gardent
CNRS / LORIA

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

Not all Data is Text

Structured Data exists in many forms

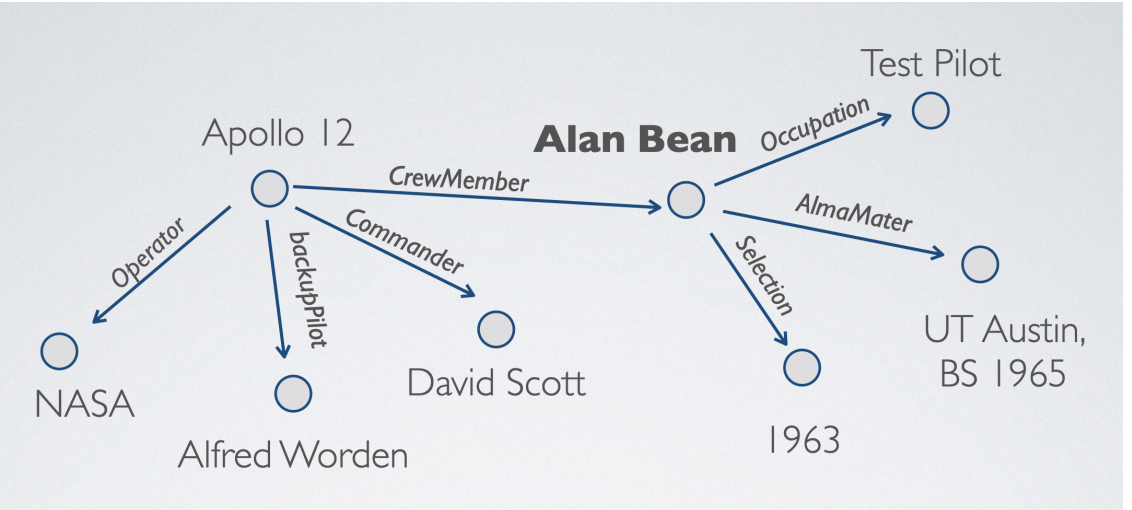
- Tabular data (e.g., csv files)
- Databases
- Records (e.g., json files)
- Knowledge Graphs (e.g., Wikidata)

Natural Language Interfaces to Structured data

- permit ***querying, interpreting or validating*** structured data
- Applications
  - Question Answering
  - Fact Verification
  - Summarisation
  - Verbalisation

Verbalising Knowledge Graphs

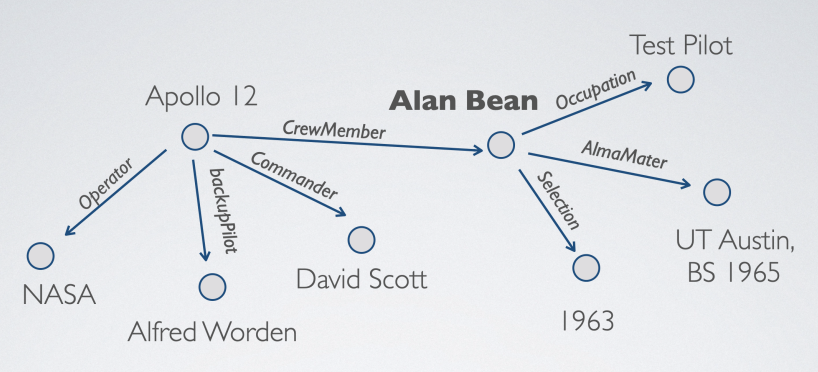
RDF Graph



English Text

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-to-Text Generation



[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。

Challenges for Multilingual KG Verbalisation

- From Structured data to Non Structured Data (Text)

Challenges for Multilingual KG Verbalisation

- From Structured data to Non Structured Data (Text)
- Structured input is underspecified

Challenges for Multilingual KG Verbalisation

- From Structured data to Non Structured Data (Text)
- Structured input is underspecified
- Lack of multilingual parallel KG/Text data

Challenges for Multilingual KG Verbalisation

- From Structured data to Non Structured Data (Text)
- Structured input is underspecified
- Lack of multilingual parallel KG/Text data
- Lack of multilingual KG/Text evaluation metrics

Challenges for Multilingual KG Verbalisation

- From Structured data to Non Structured Data (Text)
- Structured input is underspecified
- Lack of multilingual parallel KG/Text data
- Lack of multilingual KG/Text evaluation metrics
- Decoding into languages with varied morphology and word order

Outline

- The WebNLG Challenge

Outline

- The WebNLG Challenge
- Models

Outline

- The WebNLG Challenge
- Models
- Evaluation

Outline

- The WebNLG Challenge
- Models
- Evaluation
- Generalisation to Out of Domain Data

The WebNLG Challenge

Claire Gardent, Anastasia Shimorina, Shashi Narayan and Laura Perez-Beltrachini. Creating Training Corpora for NLG Micro-Planners. ACL 2017.

The WebNLG Challenge

2017: Verbalising KGs into English (KG-to-Text Generation).

2020: Verbalising KGs into English and Russian;
Converting Text to KGs (Semantic Parsing).

2023: Verbalising KGs into Low Resource Languages
(Maltese, Breton, Irish, Welsh)

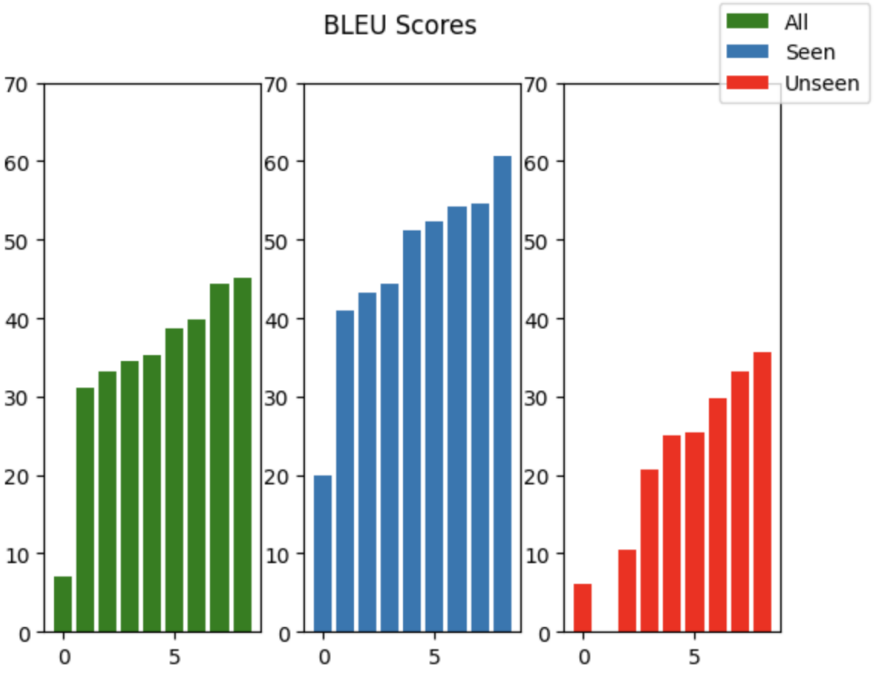
WebNLG 2017: KG ⇒ English

- Seen and Unseen Data to test generalisation
  - DBPedia graphs describing entities of various categories.
  - 10 **seen** categories: Astronaut, University, Monument, Building, Comics Character, Food, Airport, SportsTeam, City and WrittenWork
  - 5 **unseen** categories: Athlete, Artist, MeanOfTransportation, CelestialBody, Politician
- English texts are crowdsourced
- Reference Based Evaluation metrics: BLEU, METEOR etc.

	Train+Dev	Test (Seen Category)	Test (Unseen Category)	TOTAL
# (Graph,Text)	20,370	2,495	2,413	**25,298**
# Graphs	7,812	971	891	**9,674**

Gardent et al. ACL 2017. Creating Training Corpora for NLG Micro-Planners.

WebNLG 2017: KG $\Rightarrow$ English



Models: 3 Rule-Based, 1 SMT, 5 Neural

Strong differences between models

All models degrades on Unseen Data

WebNLG 2020

KG $\Leftrightarrow$ NL

Russian and English

Castro-Ferreira et al. 2020. The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task: Overview and Evaluation Results.

WebNLG 2020: KG $\Rightarrow$ English

	Train	Dev	Test NLG/SP	TOTAL
# (Graph,Text)	35,426	4,664	5,150	47,395
# Graphs	13,211	1,667	1,779	17,409

More Training Data

- 16 categories: Astronaut, University, Monument, Building, Comics Character, Food, Airport, SportsTeam, City, WrittenWork, Athlete, Artist, CelestialBody, MeanOfTransportation, Politician, Company

More Unseen Test Data

- 3 **unseen** categories: Film, Scientist, and MusicalWork
- **Unseen entities**: graphs from seen categories, but describing an unseen entity. E.g., *Nie Haisheng* in category *Astronaut*

WebNLG 2020: Participation

17 teams submitted 48 system runs.

Industry

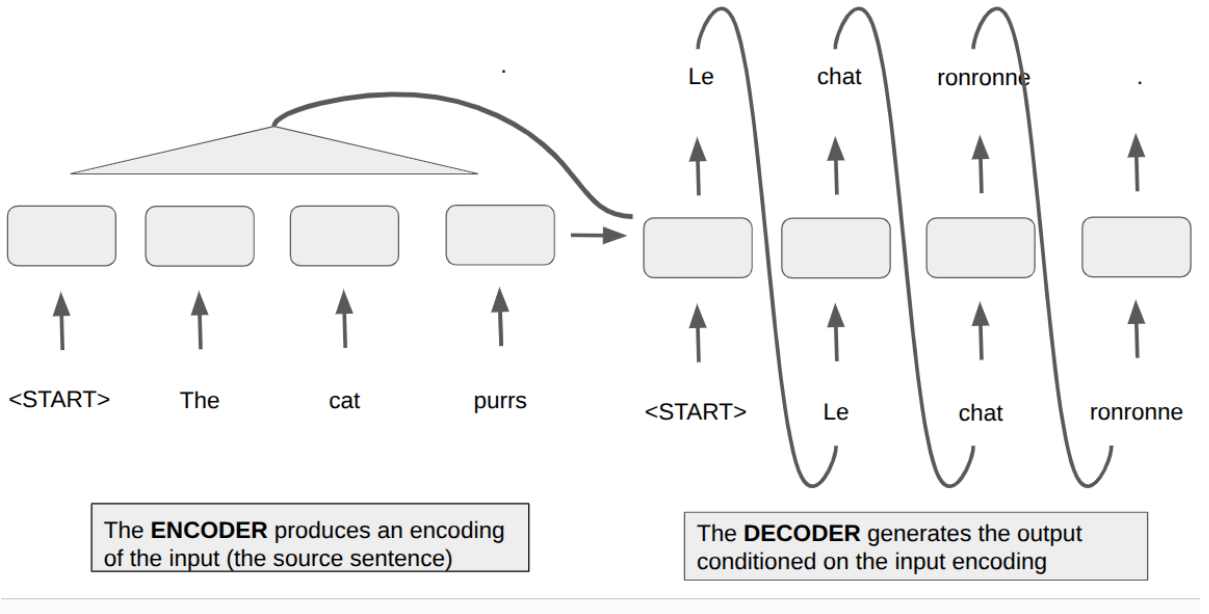
- Google, Amazon AI, Facebook, Huawei, Sber AI Lab, Orange Labs, AIST

Academia

- Montreal, Prague, Barcelona, Sao Paulo, Dublin, Ohio State

The best systems fine-tune pre-trained Encoder-Decoders (T5, BART) on the WebNLG training data.

The Encoder-Decoder Architecture



An ***encoder-decoder*** model

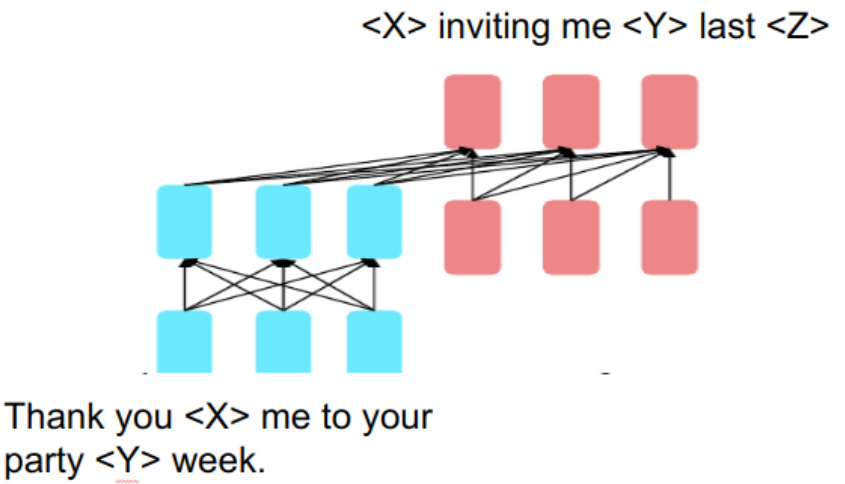
- consists of *two neural networks*
- the *encoder* creates a numerical representation of the input and
- the *decoder* generates text from left to right, one word at a time based on this representation

T5 - A Pre-Trained Encoder-Decoder

Encoder-decoder Transformer (220M parameters)

Pretraining on C4 (750 GB)

- *Masked Language Modelling (MLM) Objective with **span masking***



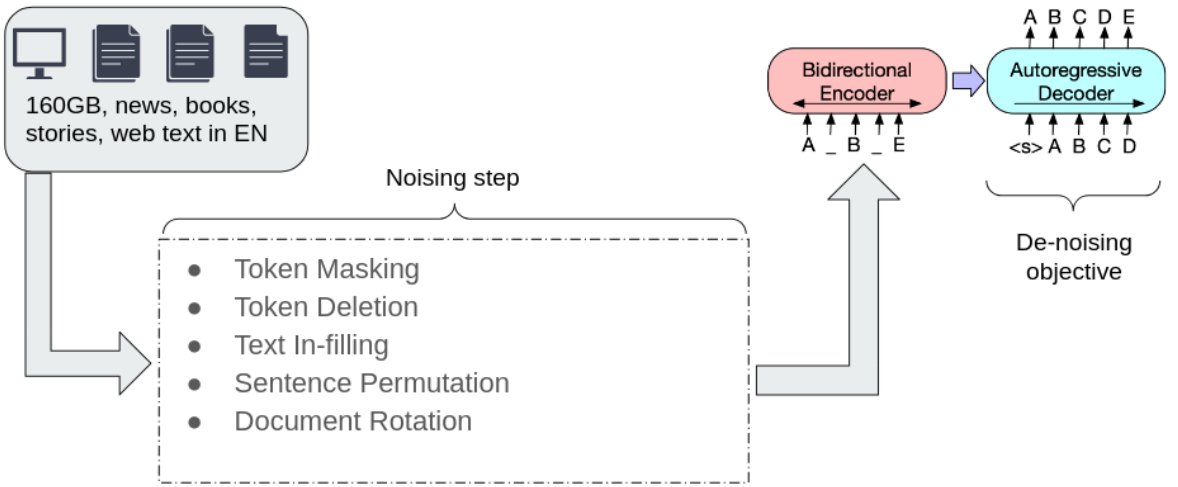
- Multi-task Learning: MT, translation, text classification ...

BART

Encoder-Decoder Transformer

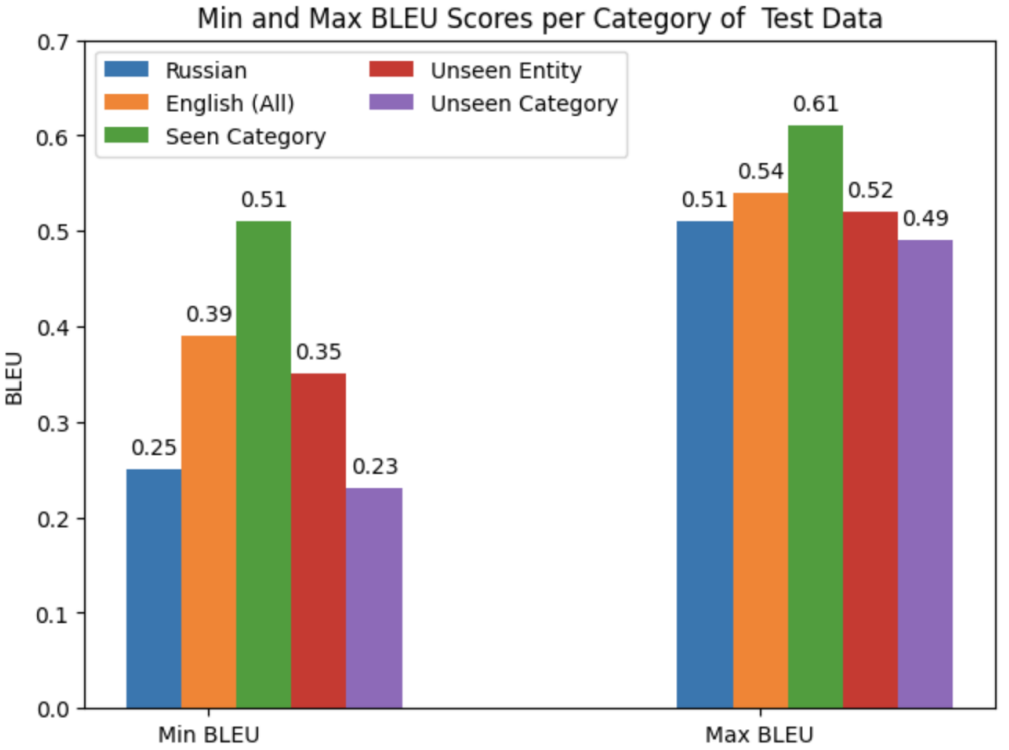
Pre-trained on 160GB of news, books, stories and web text

- *Denoising auto-encoder:* trained to ***reconstruct original text***



Lewis et al. ACL 2020

WebNLG 2020 Results for English and Russian for KG-to-Text



English >> Russian

English 2020 >> English 2017

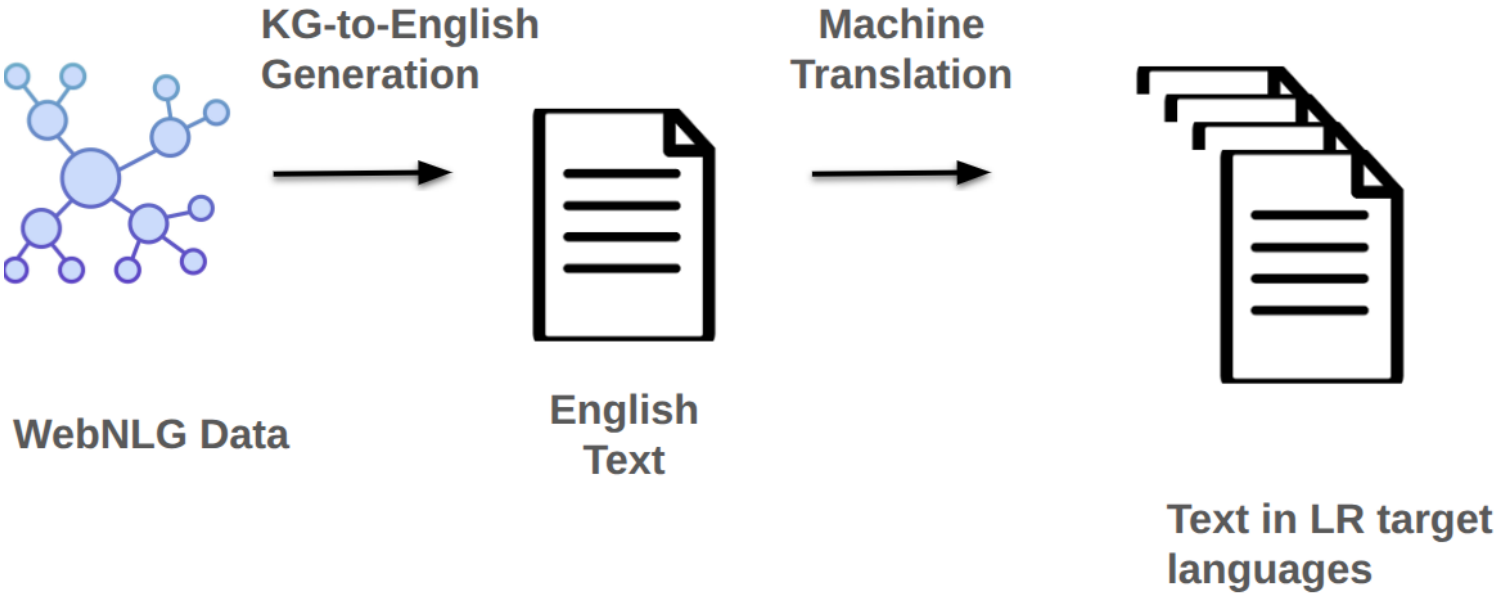
WebNLG 2023: Generating into Low Resource Languages

- Graphs from English WebNLG 2020
- Texts from professional translators
- ***No GOLD training data***

	Silver Train	Dev	Test
Breton	13,211	1,399	1,778
Welsh	13,211	1,665	1,778
Irish	13,211	1,665	1,778
Maltese	13,211	1,665	1,778

Cripwell et al. 2023. The 2023 WebNLG Shared Task on Low Resource Languages. Overview and Evaluation Results

WebNLG 2023: Pipeline NLG+MT Models



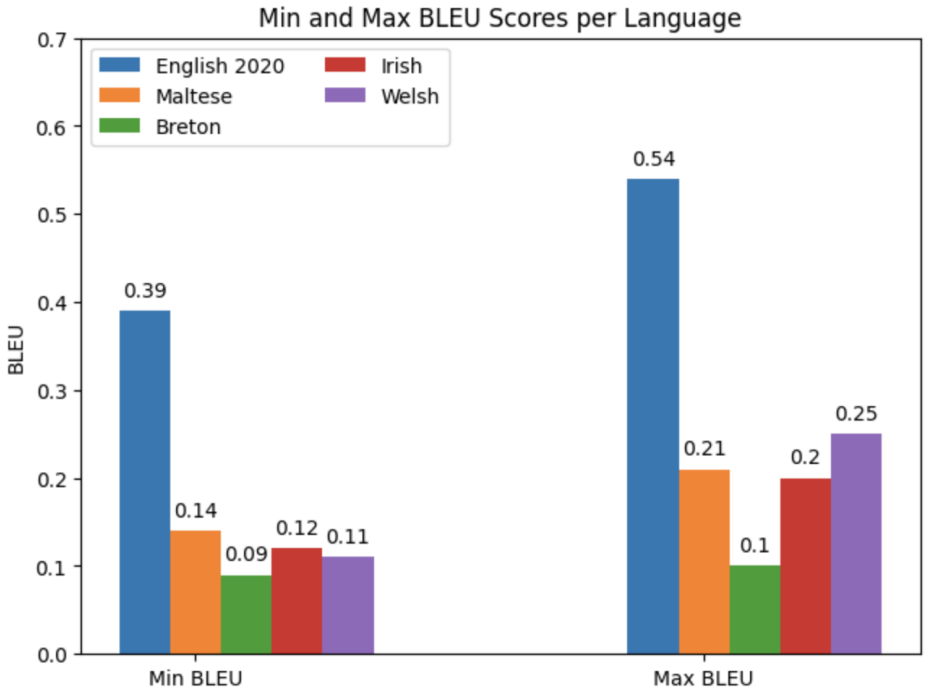
KG → English

- (m)T5 fine-tuned on English WebNLG data
- GPT3-5 in context learning

English → LR Language

- Facebook NLLB MT Model or Google Translate

WebNLG 2023 Results for KG-to-LRL Generation



Strong degradation compared to English

Very poor output for Breton

KG-to-Text for Low Resource Languages

William Soto-Martinez, Yannick Parmentier and Claire Gardent. Phylogeny-Inspired Soft Prompts For Data-to-Text Generation in Low-Resource Languages. AACL 2023.

Pipeline vs. End-to-End

For Breton, there is no (good) MT system

Pipeline vs. End-to-End

For Breton, there is no (good) MT system

 NLG+MT pipeline

Pipeline vs. End-to-End

For Breton, there is no (good) MT system

 NLG+MT pipeline

 Parameter Efficient Fine Tuning (PEFT)

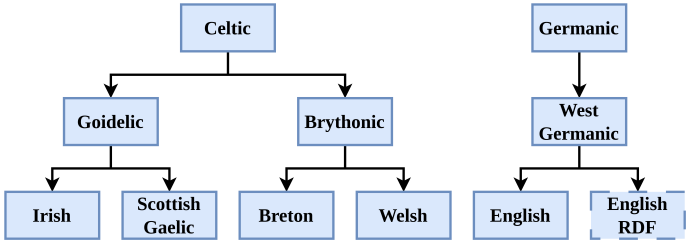
Pipeline vs. End-to-End

For Breton, there is no (good) MT system

✗ NLG+MT pipeline

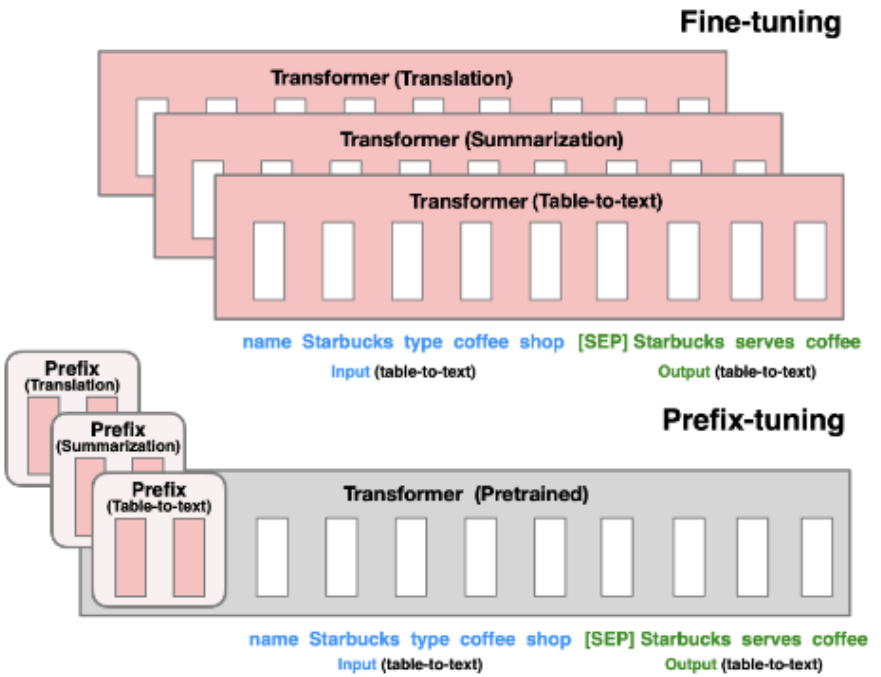
✓ Parameter Efficient Fine Tuning (PEFT)

- Soft-Prompt (Prefix Tuning)
- Structured to capture language relatedness and various tasks



Prefix Tuning (Soft Prompt)

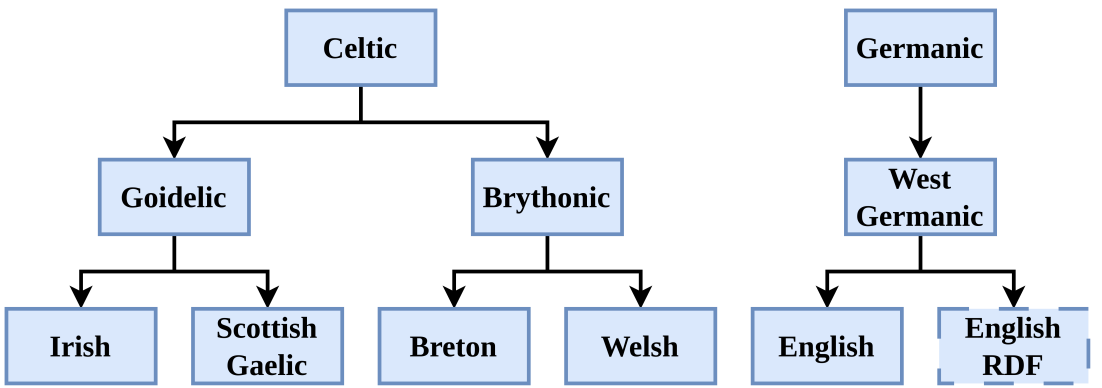
- All parameters of the pre-trained model are frozen
- Only learn the prefix (soft-prompt) parameters



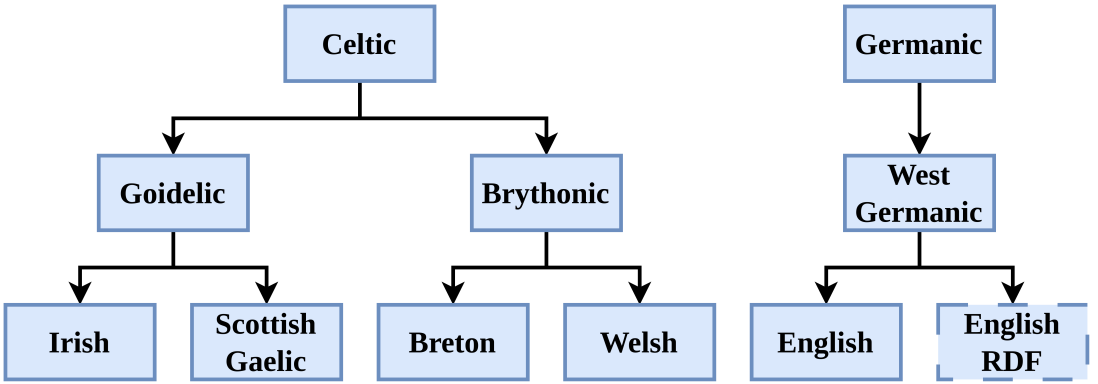
Structured Soft Prompt

The soft-prompt is decomposed into Family, Genus, and Language sub-prompts to capture the differences and similarities between languages.

50 Tokens Task	15 Tokens Source Family	15 Tokens Source Genus	15 Tokens Source Language	15 Tokens Target Family	15 Tokens Target Genus	15 Tokens Target Language	n Tokens Input Sequence



Phylogenetical Prefix Tuning



The soft-prompt structure:

Facilitates transfer: Each language benefits from the training data of their related languages.
E.g., The sub-prompt for Goidelic is updated each time an Irish or Gaelic training instance is processed.

Reduces noise: Prevents the mixture of training data to introduce too much noise to the model.

Training and Testing

Step 1: Self-supervised Training (Language Models)

Trains the Soft Prompt on unsupervised, monolingual tasks

	Task	Source			Target			Original Input Sequences					
		Family	Genus	Lang.	Family	Genus	Lang.						
Input Batch	Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
	Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
	Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
	Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
	Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Training and Testing

Step 1: Self-supervised Training (Language Models)
Trains the Soft Prompt on unsupervised, monolingual tasks

Input Batch	Task	Source			Target			Original Input Sequences					
		Family	Genus	Lang.	Family	Genus	Lang.						
	Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
	Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
	Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
	Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
	Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Step 2: Fine-Tuning on Dev KG-to-Text data (KG-to-Text Models)
Trains the KG-to-Text Task sub-prompt for each target language

Training and Testing

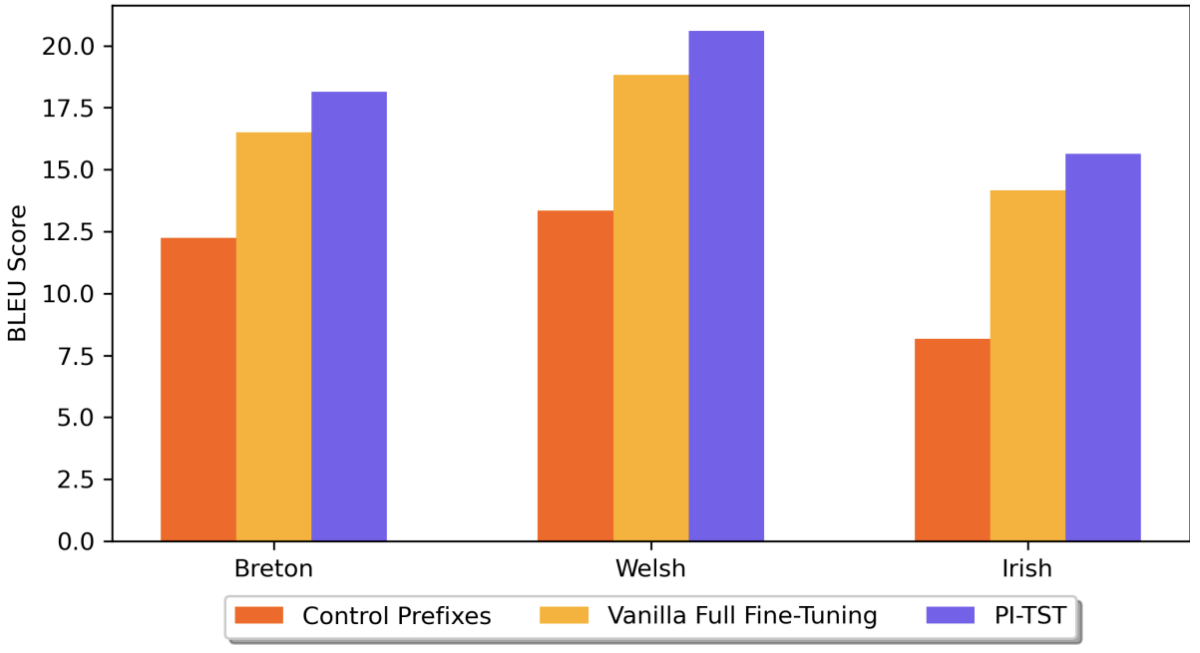
Step 1: Self-supervised Training (Language Models)
Trains the Soft Prompt on unsupervised, monolingual tasks

Input Batch	Task	Source			Target			Original Input Sequences					
		Family	Genus	Lang.	Family	Genus	Lang.						
	Masked LM	Germanic	West Germanic	RDF	Germanic	West Germanic	RDF	<S>	Einstein	<P>	<mask>	<P>	Poland
	Prefix LM	Germanic	West Germanic	English	Germanic	West Germanic	English	Thank	you	for	<mask>	<pad>	<pad>
	Suffix LM	Celtic	Britonic	Welsh	Celtic	Britonic	Welsh	<mask>	honno	?	<pad>	<pad>	<pad>
	Deshuffling	Celtc	Britonic	Breton	Celtic	Britonic	Breton	skuizh	?	out	Ha	<pad>	<pad>
	Generate	Celtc	Goidelic	Irish	Celtic	Goidelic	Irish	Seo	<mask>	<pad>	<pad>	<pad>	<pad>

Step 2: Fine-Tuning on Dev KG-to-Text data (KG-to-Text Models)
Trains the KG-to-Text Task sub-prompt for each target language

Inference
The Language sub-prompt is set to the target language.

Results



Phylogenetic prefix-tuning outperforms full fine-tuning and a SoTA approach for KG-to-Text generation

Breton: 10 (NLG+MT) $\rightarrow$ 18.15 BLEU (PEFT)

KG-to-Text for High- and Low-Resource Languages

Yifei Song, William Soto-Martinez, Anna Nikiforovskaja, Evan Chapple and Claire Gardent. Multilingual Verbalisation of Knowledge Graphs. EMNLP Findings 2025.

Three Methods, 9 Languages

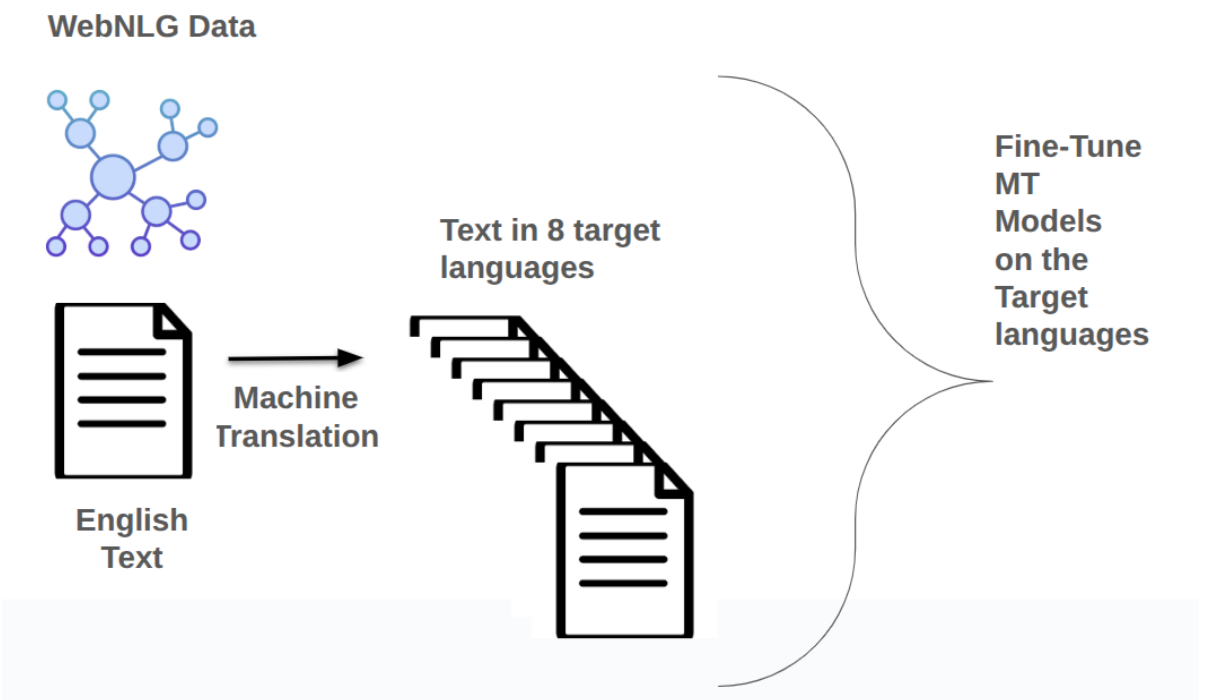
Methods

- FMT: Fine-tuning Machine Translation Models on KG/Text data
- NLG+MT: Generate into English and machine translate into the target languages
- LLM Prompting

Languages

- HRL: Chinese, French, Russian, Spanish, English
- LRL: Breton, Irish, Maltese, Welsh

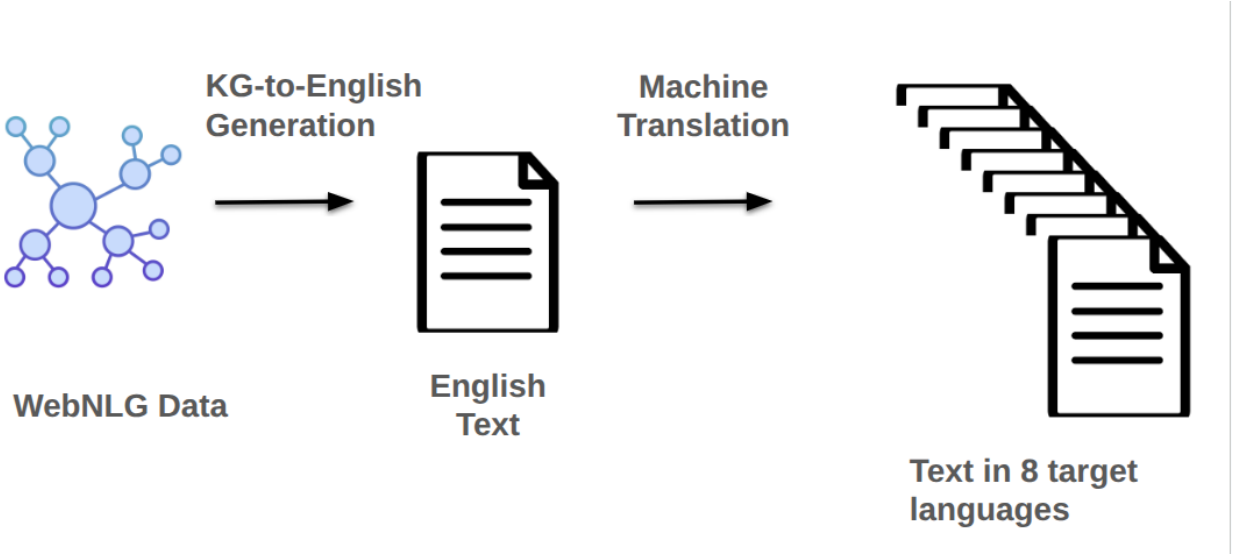
Method 1: Fine-Tuned Machine Translation



We select the best MT Model for each target language

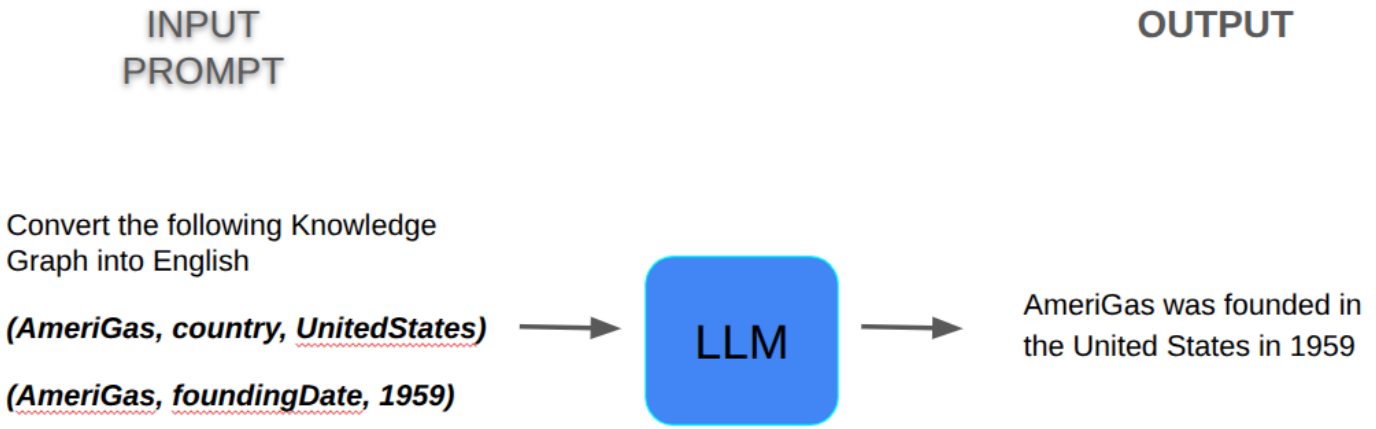
Language	H	NLLB	mBART	M2M
Chinese				✓
French	✓			
Russian				✓
Spanish	✓			
Breton	✓			
Irish		✓		
Maltese	✓			
Welsh				✓

Method 2: NLG+MT Pipeline



A two steps pipeline where a KG-to-Text model is used to generate English text from the input graph and then an MT model is applied to generate text in the target languages.

Method 3: Few-Shot Prompting

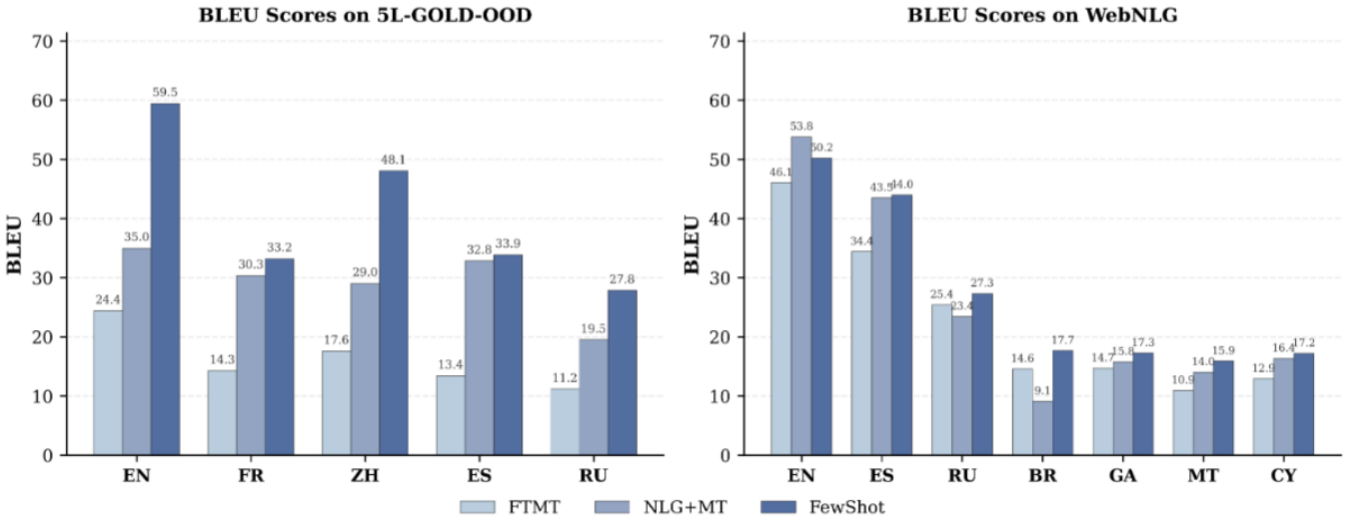


Method 3: Prompting

Prompt	Generate a lexicalisation of the following graph in French. Include all the information from the triples regardless of their type or relevance.
Graph	("subject": "AmeriGas", "property": "country", "object": "United States")
Labels	"country": "pays", "United States": "États-Unis"
Text	"AmeriGas est situé aux États-Unis."
Graph	("Western Goals Foundation", "property": "country","object": "United States")
Labels	"country": "pays", "United States": "États-Unis"
Text	

- 3-shot prompting
- we also include in the prompt, the target language labels of the entities and properties present in the input graph.
- Models: Llama-3.1-8B-instruct, Llama-3.1-70B-instruct, Qwen2.5-7B-instruct, and Qwen2.5-72B-instruct.

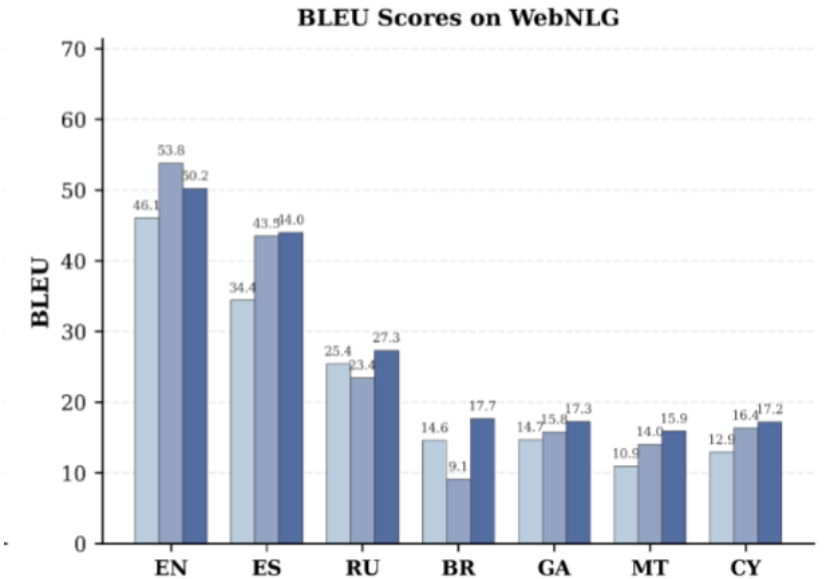
Results



On Out-Of-Domain data (Left):

- Prompting outperforms the other two methods
- The improvement mirrors the language distribution of the LLMs' pre-training corpora. The FewShot LLM has a strong advantage on languages it was well-trained on (English, Chinese), whereas in languages less represented in the LLM’s training data, the gap between FewShot and the other methods is smaller.

Results



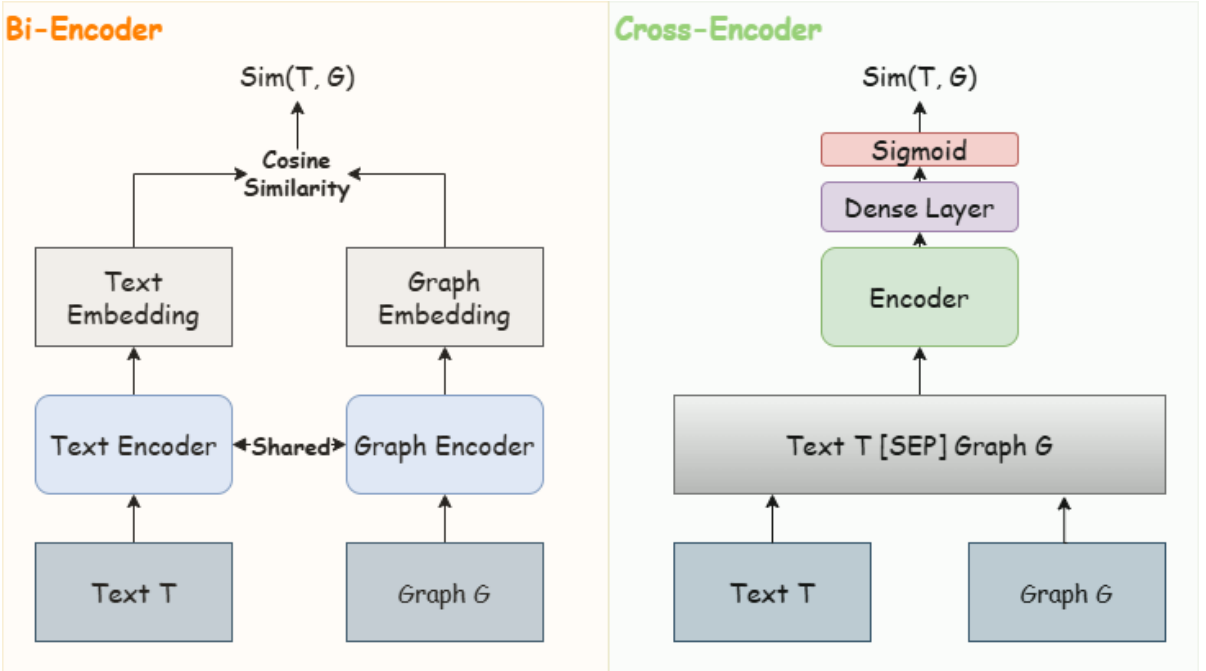
On In-Domain data :

- Prompting is best except for English
- The improvement is largest for LRL demonstrating the ability of LLMs to generate multilingual text even for a task with minimal or no training data.

MuCAL: A multilingual KG/Text Similarity Metrics

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

Learning a KG/Text Similarity Metrics



The models are trained to predict a similarity score for KG/Text pairs

Multilingual data: each graph is associated with verbalisation in 5 languages (English, German, French, Chinese, Spanish)

Contrastive Loss Training

Positive examples: all 6 verbalisations of a graph

Negative examples: Mismatched KG/Text pairs

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

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

We evaluate using 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

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

Our Model improves on previous work

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

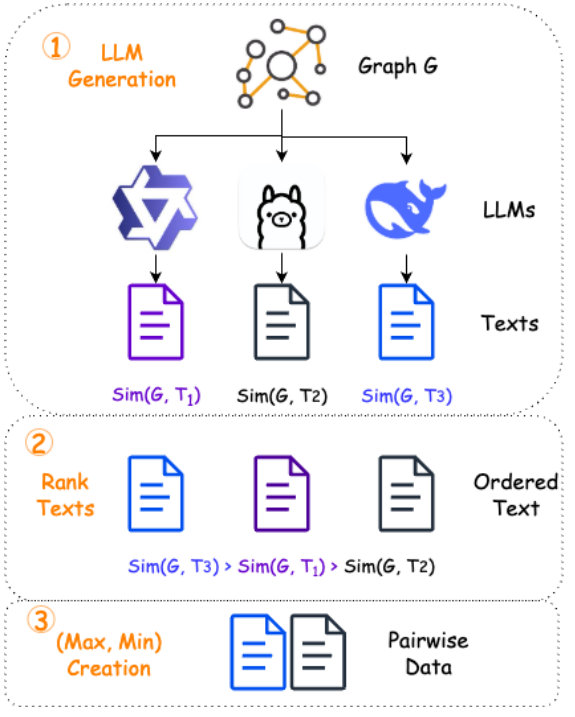
(KG, good text, bad text)

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 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
- Fine tuned on Kelm-Q1

Evaluation

Test Sets

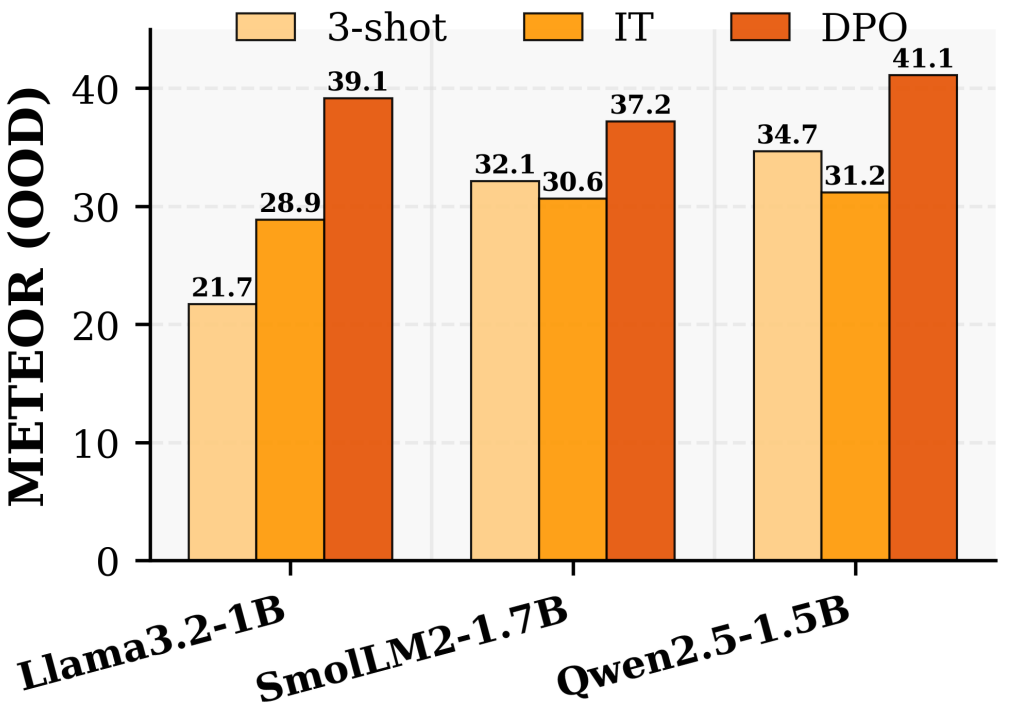
- **In-domain:** KELM-Test
- **Out-of-Domain but Public:** WebNLG
- **Out-of-Domain:** GOLD-OOD-50

Metrics:

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

DPO models generalise better to OOD data

DPO outperforms Instruction Tuning and 3-Shot Prompting



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

Perspectives

Open Questions for KG Verbalisation

- Unsupervised or Self-supervised verbalisation into Low and Medium Resource Languages
- Multiligual, reference-less Evaluation Metrics
- Multilingual Structured Prediction
- Generic models for the verbalisation of structured data (Tabular data, Attribute-Value Data, Knowledge Graphs)

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