JOINT REPRESENTATIONS OF TEXT AND KNOWLEDGE GRAPHS FOR RETRIEVAL AND EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

A key feature of neural models is that they can produce semantic vector representations of objects (texts, images, speech, etc.) ensuring that similar objects are close to each other in the vector space. While much work has focused on learning text, image, knowledge-base (KB) and image-text representations, there are no aligned cross-modal text-KB representations. One challenge for learning such representations is the lack of parallel data. We train retrieval models on datasets of (graph, text) pairs where the graph is a KB subgraph and the text has been heuristically aligned with the graph. When performing retrieval on WEBNLG, a clean parallel corpus, our best model achieves 80% accuracy and 99% recall@10, showing that similar texts and KB graphs are mapped close to each other. We use this property to create a similarity metric between English text and KB graphs, matching state-of-the-art metrics in terms of correlation with human judgments even though, unlike them, it does not require a reference text to compare against.

1 INTRODUCTION

Neural approaches have progressed in capturing semantic relatedness between larger and larger text units, from Word2Vec (Mikolov et al., 2013) to SBERT Reimers & Gurevych (2019). Such models were shown to perform well on a wide array of semantic similarity tasks, helped in part by dense retrieval systems like DPR (Karpukhin et al., 2020a).

Other work has shown that deep representations of knowledge bases (KBs) help improve such tasks as few shot link prediction, analogical reasoning Pezeshkpour et al. (2018); Pahuja et al. (2021), entity linking Yu et al. (2020) or cross-lingual entity alignment Chen et al. (2018); Xu et al. (2019).

In this work, we focus on learning cross-modal representations for English text and KB graphs which allows us both to leverage strong existing pre-trained models and to interface with text data. We consider KB graphs in RDF (Resource Description Framework, Miller (1998)) format, a semantic web standard where graphs are sets of *(subject, predicate, object)* triples. Given some aligned RDF-text data, our model learns fixed-length latent representations for texts and RDF graphs such that texts and RDF graphs that are semantically similar, are close in vector space. This enables retrieval across modalities, and allows us to create a cross-modality similarity score which can be used to evaluate the output of RDF-to-text generation models.

One challenge for learning cross-modal RDF-text representations is the lack of parallel data. We train our models on various RDF-text datasets which were created using distant supervision techniques, either combining these datasets or using them in isolation. We then compare the performance of the resulting retrieval models (i) on the WEBNLG dataset, a parallel RDF-text dataset where texts are crowdsourced to match the graph (texts and graphs are semantically equivalent) and (ii) on WIKICHUNKS, a more challenging, less well aligned dataset which imitates the conditions in which retrieval on Wikipedia is usually executed. We observe marked differences between the models, which suggests differences in alignment quality between the three datasets, and we show that our models outperform a strong natural language-only baseline by a large margin.

Distance within embedding space can be used to evaluate the output of RDF-to-text generation models (Is the generated text similar to the input graph?). In order to evaluate this metric, we compute correlations between the similarity score output for a graph-text pairs by our model and human judgments of semantic adequacy (input/output semantic similarity) using ratings from the 2020

WebNLG Challenge. After fine-tuning on data from the 2017 WebNLG challenge, as well as introducing new classes of data augmentation at pre-training time, our best system is better or on par than existing metrics at correlating with human evaluation, even though it does not require a reference for comparison, as is the case for most NLG evaluation metrics such as BLEU (Papineni et al., 2002), TER (Snover et al., 2006), BLEURT (Sellam et al., 2020b), METEOR (Banerjee & Lavie, 2005) or BERT-Score (Zhang* et al., 2020).

Our contributions can be summarised as follows.

- We train a cross-modal RDF-text model to learn aligned (RDF graph, text) representations, making it suitable for cross-modal retrieval. We show that this retrieval model outperforms a state-of-the-art text-text retrieval model by a large margin, demonstrating the effectiveness of our cross-modal representation learning model against a retrieval model which represents RDF graphs using a standard text encoder. Training on various aligned datasets allows to analyze their respective quality.
- We provide a novel, evaluation metric for RDF-to-text generation models by combining bi- and cross-encoder training procedures and adding adversarial data to address the models' weaknesses. We show that this new metrics outperforms other existing RDF-to-text evaluation metrics in terms of correlation with human judgements of semantic adequacy, even though it does not require a costly human reference to compare against.

2 Related Work

We briefly review recent approaches to uni- and cross-modal retrieval, representation learning models and evaluation metrics for Natural Language Generation (NLG) models.

Natural Language Retrieval Models. For natural language, a first class of retrieval models focuses on retrieving sentences that are similar to some input sentence. BERT (Devlin et al., 2019) has been used as a cross-encoder. Two sentences are given with a separator token, cross-attention applies to all input tokens and the resulting representation is fed into a linear layer to score the match. However, this is computationally inefficient as it is not possible to pre-compute and index such representations. A pre-computable model was proposed by (Reimers & Gurevych, 2019) who used twin encoders pre-trained on Natural Language Inference data (Bowman et al., 2015) to set new state-of-the-art performance on a large set of sentence scoring tasks. Further work (Chen et al., 2020; Humeau et al., 2019) combined cross- and bi-encoders to reach a tradeoff between accuracy and efficiency. We differ from those works in that we focus on cross-modal representation learning and retrieval models.

Representation Learning for Knowledge-Bases. Various KB embedding models have been proposed to support downstream applications such as KB completion or alignment of different bases. Compositional approaches Nickel et al. (2011; 2016) use tensor products to model relations as functions of their argument entities. Translational approaches model relations as translations operations from subject (head) to object (tail) entity Bordes et al. (2013); Yang et al. (2014); Trouillon et al. (2016). Neural models have also leveraged 2-D convolutions over entity embeddings to predict relations Dettmers et al. (2018) as well as graph convolutional networks Schlichtkrull et al. (2018). All these approaches focus on representation learning for Knowledge-Bases entities and relations. In contrast, we focus on cross-modal similarity between a text and a KB graph.

Cross-Modal Representation Learning and Retrieval. Some work has focused on incorporating natural language information to improve KB representations. Han et al. (2016); Toutanova et al. (2015); Wu et al. (2016) encode words and KB entities into a single vector space, and Wang & Li (2016); Yamada et al. (2016) learn word and entity embeddings separately then map them into a shared space. Both approaches use text as additional training signal to improve KB representations, and limit themselves to word-level information. Instead, we focus on scoring the similarity between arbitrary-length natural language text and a KB graph. We are not aware of any extant such text-KB models. The best-known cross-modal contrastive model is Radford et al. (2021), which pre-trained an image-text match scoring model.

Evaluation metrics for Natural Language Generation Models. Surface-based metrics such as BLEU (Papineni et al., 2002) which measure token overlap between generated and reference text, are commonly used. Methods such as BERT-Score (Zhang* et al., 2020) or BLEURT (Sellam et al., 2020a) which leverage neural representations are currently state-of-the-art. All these methods compute a score by comparing the generated text with human-produced references, rarely available and costly to produce. Some metrics evaluate the generated output with respect to the input rather than to a reference. Wiseman et al. (2017) use the precision of input relations found in the output texts. Dušek & Kasner (2020) use a natural language inference pre-trained model to score input-output two-way entailment. For data-to-text generation specifically, Rebuffel et al. (2021) introduce Data-QuestEval, which uses question answering to compare input graph and output text.

3 LEARNING CROSS-MODAL RDF-TEXT REPRESENTATIONS

3.1 MODEL

Similar to Schroff et al. (2015); Reimers & Gurevych (2019), we use twin Transformer encoders to create RDF and text representations such that the embeddings of an RDF graph and of a piece of text with similar content are close in the vector space. A mean-pooling operation creates fixed-sized embeddings $embed(x)$ for x either an RDF graph or a text. RDF graphs are linearized as "[S] <subject$_1$> [P] <property$_1$> [O] <object$_1$> ... [S] <subject$_n$> [P] <property$_n$> [O] <object$_n$>" where "[S]", "[P]", "[O]" serve as special tokens and are added to the tokenizer vocabulary. This allows us to treat any knowledge base format.

We train this system using a contrastive loss with *in-batch negatives* (Henderson et al., 2017). This variant of contrastive loss computes the pairwise similarities between every text and every RDF in the batch. A softmax is then applied on the RDF axis, which creates a multi-class classification problem: every text data point must be matched to the parallel RDF. The loss can be written as :

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

$$sim(text_i, rdf_j) = \cos(embed(text_i), embed(rdf_j))$$

with I the set of training instances in the batch. Intuitively, this trains the encoder to learn representations that map text items closer to their RDF anchor than to other RDF graphs in the dataset.

In all our experiments, we start from `all-mpnet-base-v2`, a pre-trained sentence-MPNet (Song et al., 2020) model, in order to leverage its strong pre-trained text representations.

3.2 TRAINING DATASETS

For training, we need (g, t) pairs where g is a Wikidata RDF graph and t is a text in English whose content is similar to g. We compare three datasets, all created using distant supervision.

TeKGen. Agarwal et al. (2021) use heuristics to align triples from Wikidata to Wikipedia sentences. The TEKGEN dataset covers 1,041 Wikidata properties and consists of about 6M (graph, text) pairs where each text is a sentence.

KELM. The KELM corpus has 15M (graph, text) pairs where graphs are created based on relation co-occurrence counts i.e. frequency of alignment of two properties to the same sentence in the training data (Agarwal et al., 2021). Texts are then generated from these graphs using T5 fine-tuned on TEKGEN.

TREx. Elsahar et al. (2018) use word- and sentence-tokenization, coreference resolution, a date-time and a predicate linker, plus various RDF-text alignment methods to create TREX, a dataset aligning 11 million Wikidata triples 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

Table 1: **Training and test data for retrieval**. # (t,g): Number of graph-text pairs, # T: Number of texts, # G: Number of graphs, # P: Number of distinct properties, # E: Number of distinct entities.

4 EVALUATION SETUP

We evaluate our representations using a retrieval reformulation of the data-to-text NLG task: Given the embedding of a graph, how well can we identify the most similar text in the corpus? As our evaluation sets have 1-to-1 mappings between sources (the graphs) and targets (the texts), the retrieval performance in the opposite direction does not vary by more than 2%.

4.1 TEST DATASETS

We use two datasets for evaluation: WEBNLG Gardent et al. (2017) and WIKICHUNKS, which we create in this work.

WebNLG is a dataset of pairs where the texts were crowdsourced to match the input graph. In WEBNLG the RDF graph are from the DBpedia KB, whereas our models were trained on the Wikidata KB format. To assess the ability of our retrieval model to generalise to different KBs, we evaluate our model both on WEBNLG-DB, the original DBpedia-based dataset, and WEBNLG-WD where the DBPedia graphs have been mapped to Wikidata **?**.

WikiChunks consists of 7.3M graph-text pairs where the text is a 100-word *passage* from a Wikipedia dump and the graphs are matching Wikidata graphs. We create matching graphs by aligning all Wikidata *(s, p, o)* triples with a Wikipedia passage such that the subject s of that triple matches the entity described by the Wikipedia page from which the passage was extracted and the object o, or one of its aliases, is mentioned in that passage. Retrieving on this dataset imitates the conditions in which retrieval on Wikipedia is usually executed (Karpukhin et al., 2020b; Lewis et al., 2020). This is a challenging task as, contrary to WEBNLG, WIKICHUNKS matches are not aligned: the wikidata graph information is strictly included in the passage, which may contain much more. Several passages may also contain very similar information. To make evaluation easier, and because it is the same order of magnitude as WEBNLG, we use a subset of 30000 pairs.

Table 1 shows some statistics for all datasets.

4.2 BASELINE, EVALUATION METRICS AND VARIANTS

We use `all-mpnet-base-v2`, the state-of-the-art dense sentence embedding model that our models are training from, as a baseline. `all-mpnet-base-v2` is used for semantic similarity as our models have, but was only trained on text. It is otherwise evaluated in the same retrieval setting. We evaluate performance in terms of R@1/R@10, which is the percentage of graph for which the correct text is present in the 1/10 top-ranked texts.

5 RESULTS

5.1 GENERAL RESULTS

Models trained on all training sets outperform the baseline by a large margin on all test sets with an R@1 of 0.8 for our best model for each training/test set pair against 0.4 for the baseline (Figure 1). This demonstrates the effectiveness of our cross-modal representation learning model against a model which hasn't been adapted to RDF data. Not pictured as it cannot serve for comparison

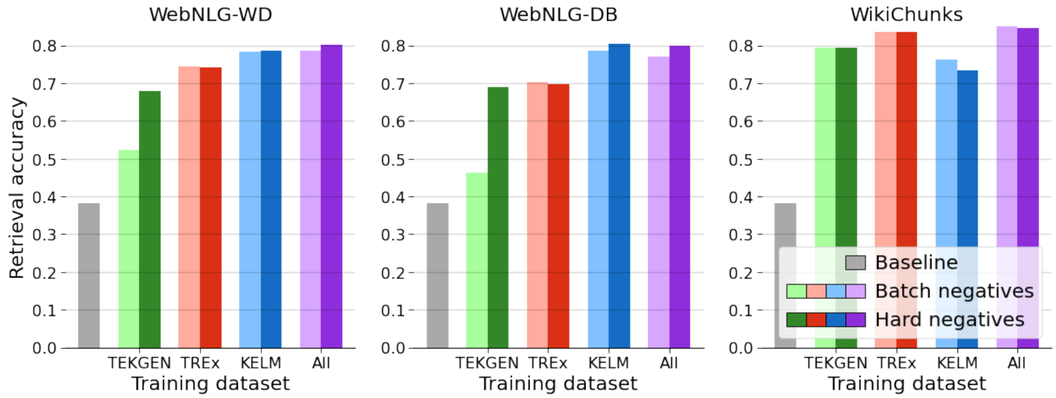

Figure 1: **Retrieval Accuracy** for a variety of training datasets and objectives. Our models outperform the baseline (left most grey bar) by a large margin. Hard negatives help across the board. Training on an equal mix of datasets yields consistently high performance on aligned (WEBNLG) and noisy (WIKICHUNKS) data.

.

is R@10, which reaches 0.98 or above for all models. **The models also generalize well across knowledge bases.** While all models are trained on Wikidata graphs, they perform similarly on WEBNLG-DB and WEBNLG-WD.

We further investigate the impact of four main factors on retrieval accuracy: batch size and types of negatives, training data quality and training data quantity.

5.2 BATCH SIZE AND NEGATIVES

We experiment with adding artificial hard negatives to the batch, and with different batch sizes. Confounders are constructed from the correct graph by corrupting a triple inside that graph, replacing a subject, object or predicate at random by another subject, object or predicate in the dataset. **This form of data augmentation is made possible by the formalized nature of RDF graphs: it would be much harder to create confounders on the text side.**

Hard vs. In-batch negatives Figure 1 shows retrieval accuracy when using only in-batch vs. using in-batch and hard negatives. We see that hard negatives mostly help when retrieving on parallel data (WEBNLG) i.e., when small graph-text mismatches strongly impact accuracy. We also see that hard negatives have the strongest impact for the model trained on TEKGEN , the model with lowest retrieval accuracy. This suggests that hard negatives are most helpful in improving retrieval when the training data is noisier than the evaluation data.

Batch size. As previous work has found that larger batch sizes improve contrastive training (Qu et al., 2021), we experiment with two batch size set-ups: 192[1] and 2560[2]. We do not find that larger batch sizes consistently improve retrieval accuracy, and keep the smaller ones for practical reasons. Figure 7 in appendix A shows detailed results.

5.3 TRAINING DATA QUALITY

The quality of training data has a strong impact on retrieval accuracy. We see that performance varies with the training data used: on WEBNLG retrieval, KELM yields by far the best results followed successively by TREX and TEKGEN. On WIKICHUNKS, which is more loosely aligned, TREX is the best dataset and KELM is slightly behind. We create an equal-mixture dataset by concatenating subsets of equal sizes of each dataset[3]. As the rightmost column in figure 1 shows, this allows

[1]The maximum we could fit on a 8-A100 cloud instance.
[2]The maximum we could fit on a larger cluster.
[3]This makes it thrice the size of the smallest dataset, TREX.

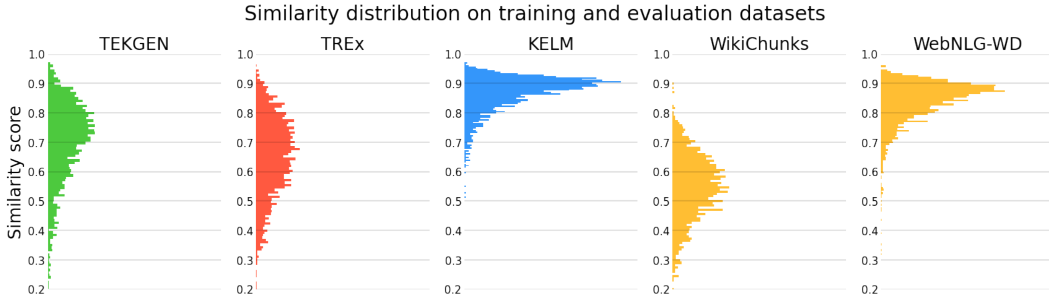

Figure 2: **Pair similarity distributions** according to `all_datasets_hard_negatives`

us to capture the best of both worlds. We dub the model trained on this data with hard negatives `all_datasets_hard_negatives`.

The similarity distributions according to `all_datasets_hard_negatives` is shown in Figure 2, which matches those results: KELM is much better aligned. This is in line with intuition as KELM text is generated from the input graphs while TREx and TEKGEN are created using distant supervision. We attempted to bootstrap dataset quality by re-training models on the 50% of the data identified as highest-similarity. We find that this does not increase performance and can sometimes even decrease it, probably because of loss of diversity.

5.4 Training Data Quantity

As shown in Figure 3, retrieval performance plateaus early in training. The advantage of KELM or the concatenated dataset is not due to their larger size.

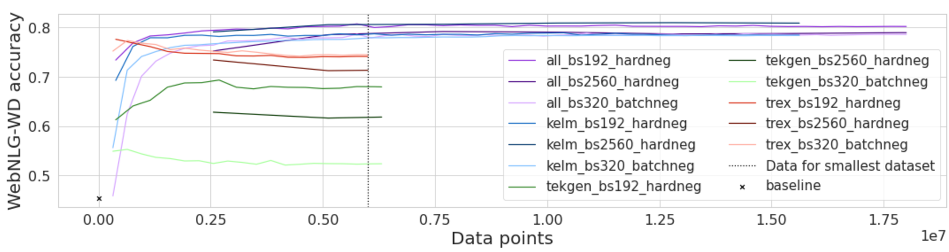

Figure 3: **Performance throughout training** evaluated by WEBNLG-WD accuracy. Training for longer than the size of the smallest datasets does not change performance meaningfully. **Larger datasets do not have an edge over smaller ones.**

6 Building a Referenceless Metric for Data-to-text Generation

Commonly-used metrics for Natural Language Generation require references to compare the output against, which must be produced by human annotators. Can we leverage our joint embeddings to compare the output to the input, reducing the necessary resources?

6.1 Learning from Human Judgments of Semantic Adequacy

Our retrieval models can be used to provide a similarity metric between text and formal data in the form of the scalar product or cosine distance in embedding space. We can further improve this metric by fine-tuning on human judgments of RDF-text adequacy. In order to show the generalization strength of this approach, we fine-tune our `all_datasets_hard_negatives` model on human-rated WEBNLG-2017 items, and evaluate on human-rated WEBNLG-2020 items, which uses different test data and different criteria for the assessment of semantic adequacy by human judges.

Shimorina et al. (2018) provides human judgments for the output of 10 NLG systems from WEBNLG challenge 2017. Each model was evaluated on a sample of 223 texts yielding a total of 2230 generated texts annotated with human judgments for the following three criteria.

- **Semantic adequacy**: Does the text correctly represent the meaning in the data?
- **Grammaticality**: Is the text grammatical (no spelling or grammatical errors)?
- **Fluency**: Does the text sound fluent and natural?

Castro Ferreira et al. (2020) provides human judgments for the output of 16 NLG systems from WEBNLG Challenge 2020. Each model was evaluated on a sample of 178 texts yielding a total of 2,848 generated texts annotated with human judgments for the following five criteria.

- **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, well-structured, written in acceptable English?
- **Fluency**: Does the text progress naturally and form a coherent, easy-to-understand whole?

We train on the 2017 *semantic adequacy* metric. To assess how well our similarity metrics reflects human judgements of similarity between an RDF graph and a Natural Language Text, we compute correlations between our systems scores and the 2020 human judgments that correspond to semantic adequacy, namely *data coverage*, *relevance*, and *correctness*[4].

6.2 FINE-TUNING PROCEDURE

Bi- and Cross-encoder ensembling We can fine-tune our pre-trained model as a *cross-encoder*, where there is only one instance of the model, which can attend to both items simultaneously and feed into a linear layer, rather than a *bi-encoder* as previously, where two instances of the model embed the two items separately and the dot product or cosine distance serves as the output. The cross-attention feature allows for higher performance at the cost of making retrieval prohibitively expensive as all n^2 distances must be computed separately Humeau et al. (2019). However, bi-encoders and cross-encoders perform well on different data points. The scores they give WEBNLG-2020 candidates have surprisingly low Pearson correlation, 0.66. This makes them good candidates for ensembling, and indeed, taking the mean of the bi- and cross-encoder scores yields higher correlation with all human judgments. Both architectures, as well as the ensembling method, are represented in diagram 4.

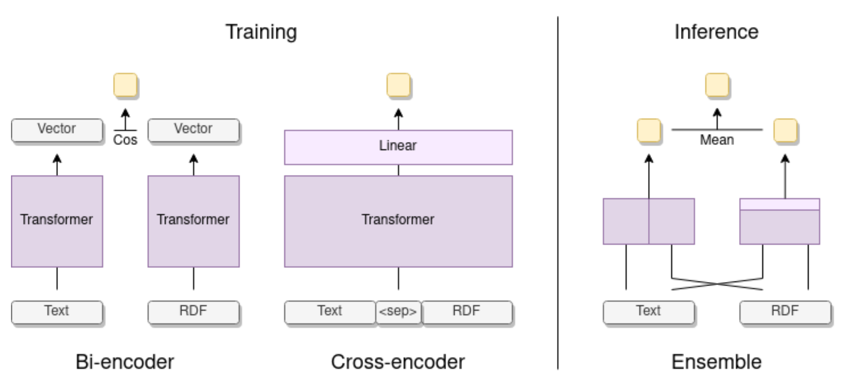

Figure 4: **Fine-tuning setup.** We fine-tune both bi-encoders and cross-encoders on human-rated data. At inference time, we use the mean of a bi-encoder and a cross-encoder as the final metric.

[4]We train on WEBNLG-2017 and evaluate on WEBNLG-2020 as semantic adequacy is a more global criteria encompassing coverage, relevance and correctness while the reverse is not true.

Robustness to inversion Transformer-based models can sometimes behave as advanced bag-of-word models (Sinha et al., 2021), which would not see a difference if the subject and object are reversed in a triple. In order to examine the robustness of our models to this behaviour, we create an adversarial dataset from all the 1-triple graphs in WEBNLG 2020 with non-symmetrical[5] relationships. In this dataset, for each text, there is a pair with the correct triple and a pair in which the triple's predicate arguments (subject and object) have been inverted e.g., *(André the Giant, larger than, Samuel Beckett)* vs. *(Samuel Beckett, larger than, André the Giant)*. This dataset (WEBNLG-INV) consists of 2793 (g, t), and (g_inv, t) pairs where (g, t) is a graph of size one with a non-symmetrical relationship in WEBNLG-WD, t is the corresponding text and g_inv is the corrupted triple.

When evaluating on this dataset, we report the difference in similarity between text and correct graph on the one had and text and corrupted graph on the other: $sim(g, t) - sim(g_{inv}, t)$. The higher the distribution is, the better the model is at recognizing predicate inversion. Figure 5 shows the results. `all_datasets_hard_negatives`, the retrieval model presented in Section 3.1, does not do well at this task, with 38% of the inverted triplets estimated more similar to the text than the original ones. (After fine-tuning on WEBNLG-2017 judgments, 30%)

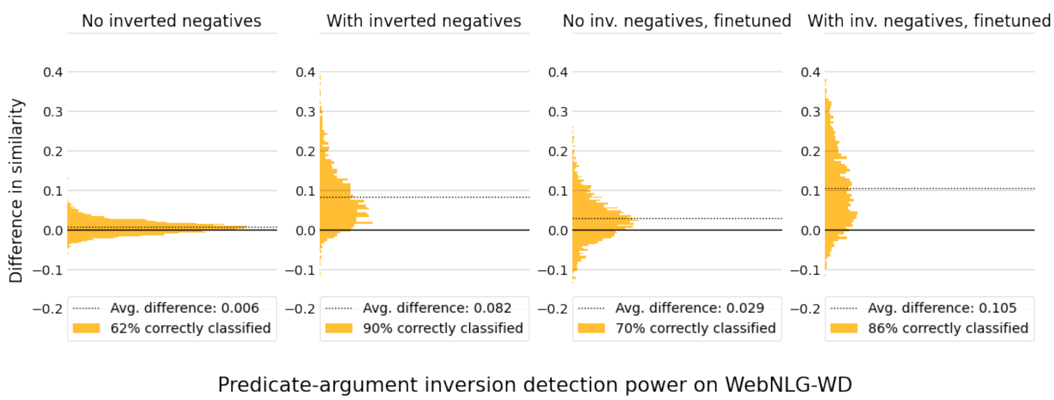

Figure 5: **Difference in similarity between correct and corrupted graph-text pairs.** On the left, `all_datasets_hard_negatives` and `all_datasets_hardinv_negatives` just after pre-training, and on the right, both models after fine-tuning and ensembling on WEBNLG-2017. The system we used as a final metric is the last plot on the right. Models that have seen inverted negatives at pre-training can better distinguish between correct and corrupted pairs.

In order to make our models robust to inversion, at pre-training time, we add inverted negatives to the mix of artificial negatives in the batches: confounding graphs where a random triplet has been inverted. The resulting model, `all_datasets_hardinv_negatives` has the same retrieval accuracy, but gains inversion detection abilities. This ability is conserved through fine-tuning, as Figure 5 shows: only 14% of triplets are misclassified.

The final system we choose as a metric is the ensemble of a bi- and cross-encoder pre-trained on the concatenation of KELM, TEKGEN and TREX using contrastive learning with our two types of data augmentation, then fine-tuned on WEBNLG-2017 human judgments.

6.3 COMPARISON WITH OTHER EVALUATION METRICS

Correlation with human judgments are shown in Figure 6 for a variety of automated evaluation metrics: three metrics that require a reference (BLEU, BERTscore-F1, BLEURT) and two referenceless metrics (Data-QuestEval and our ensembled metric). We present Pearson and Spearman correlations for the sake of completeness. On both of them, our metric is the best-performing referenceless metric. It has better (+0.053 on the average of the human judgments) Pearson correlations and worse (-0.047 on average) Spearman correlations than BLEURT, the previous best-performing metric, making them about matched. Scatter plots of the underlying distributions are given in figure 8 in appendix B.

[5]Manually defined. The list is in appendix C.

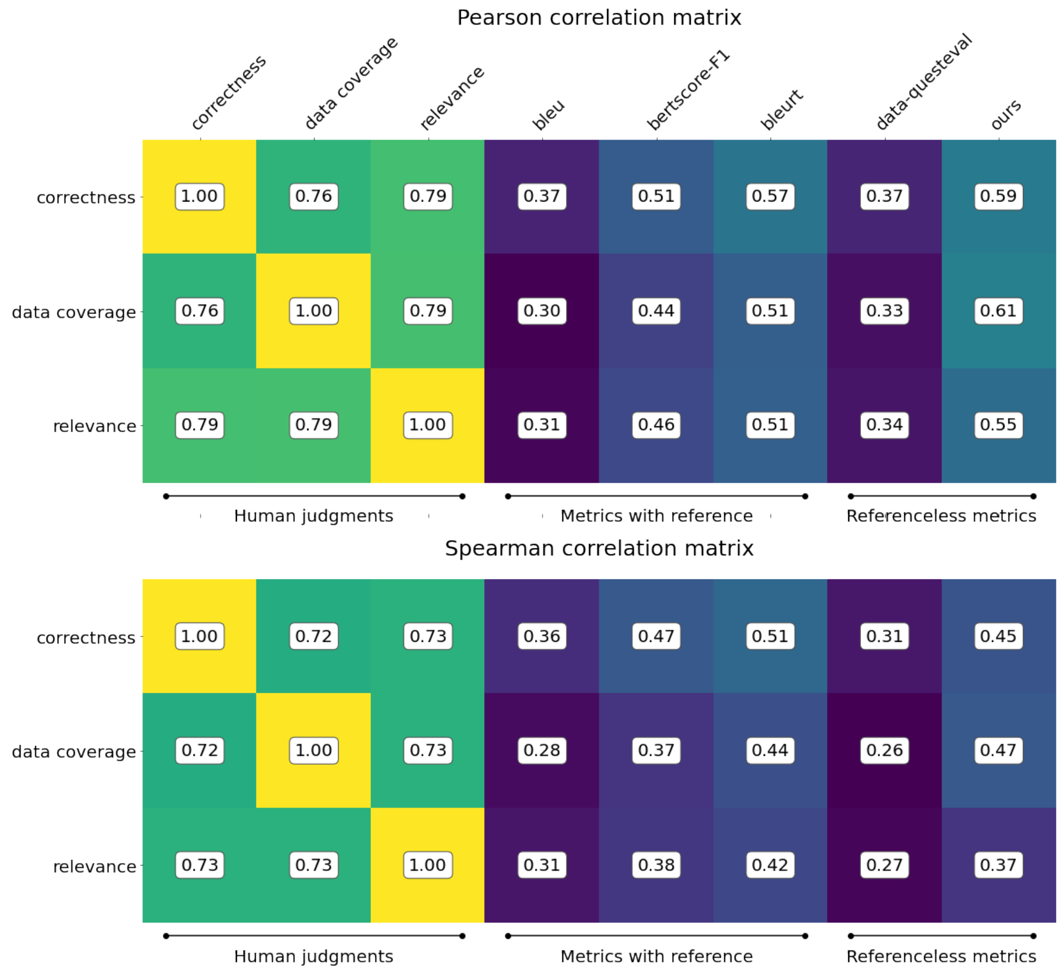

Figure 6: **Pearson and Spearman correlation between automatic metrics and human judgments.** Lighter and higher is better. Our metric outperforms the other referenceless metric and matches BLEURT, which requires a reference.

As human references are rarely available and costly to produce, and our metric attains the same level of correlation with human judgments without relying on them, it is the most practical choice to evaluate data-to-text generation. In this case, it was not fine-tuned to the same kind of data it was applied to, showing its generalization performance to new datasets. If one has a specific dataset or task in mind, even better performance could be attained by training on a set of specific of human judgments.

7 Conclusion

We presented an architecture and pre-training strategy to measure the similarity between RDF graphs and English texts, introducing novel data augmentation strategies made possible by the RDF structure. Specifically, we introduced a bi-encoder retrieval model trained on unlabeled RDF-text data which achieves high retrieval accuracy on both parallel and real-life, less well aligned datasets. Building from this pre-trained model, we further provided a novel evaluation metric for RDF-to-text generation models which matches state-of-the art reference-using metrics and outperforms existing reference-less metrics in terms of correlation with human judgments of semantic adequacy.

REFERENCES

Oshin Agarwal, Heming Ge, Siamak Shakeri, and Rami Al-Rfou. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3554–3565, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.278. URL `https://aclanthology.org/2021.naacl-main.278`.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0909`.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075`.

Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pp. 55–76, Dublin, Ireland (Virtual), 12 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.webnlg-1.7`.

Jiecao Chen, Liu Yang, Karthik Raman, Michael Bendersky, Jung-Jung Yeh, Yun Zhou, Marc Najork, Danyang Cai, and Ehsan Emadzadeh. DiPair: Fast and accurate distillation for trillion-scale text matching and pair modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2925–2937, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.264. URL `https://aclanthology.org/2020.findings-emnlp.264`.

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 3998–4004. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/556. URL `https://doi.org/10.24963/ijcai.2018/556`.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Ondřej Dušek and Zdeněk Kasner. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 131–137, Dublin, Ireland, December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.inlg-1.19`.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1544`.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 179–188, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1017. URL `https://aclanthology.org/P17-1017`.

booktitle = "Proceedings of LREC", url=http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.29)Han, Ferreira, and Gardent]hangenerating Kelvin Han, Thiago Castro Ferreira, and Claire Gardent. Generating questions from wikidata triples. 2022], booktitle = "Proceedings of LREC", url=http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.29.

Xu Han, Zhiyuan Liu, and Maosong Sun. Joint representation learning of text and knowledge for knowledge graph completion. *ArXiv*, abs/1611.04125, 2016.

Matthew L. Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *CoRR*, abs/1705.00652, 2017. URL `http://arxiv.org/abs/1705.00652`.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Real-time inference in multi-sentence tasks with deep pretrained transformers. *CoRR*, abs/1905.01969, 2019. URL `http://arxiv.org/abs/1905.01969`.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL `https://aclanthology.org/2020.emnlp-main.550`.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. *CoRR*, abs/2004.04906, 2020b. URL `https://arxiv.org/abs/2004.04906`.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *CoRR*, abs/2005.11401, 2020. URL `https://arxiv.org/abs/2005.11401`.

Tomaš Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013. URL `https://arxiv.org/abs/1301.3781`.

Eric Miller. An introduction to the resource description framework. *D-lib Magazine*, 1998.

Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 809–816, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 1955–1961. AAAI Press, 2016.

Vardaan Pahuja, Yu Gu, Wenhu Chen, Mehdi Bahrami, Lei Liu, Wei-Peng Chen, and Yu Su. A systematic investigation of KB-text embedding alignment at scale. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1764–1774, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.139. URL `https://aclanthology.org/2021.acl-long.139`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. Embedding multimodal relational data for knowledge base completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3208–3218, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1359. URL https://aclanthology.org/D18-1359.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.466. URL https://aclanthology.org/2021.naacl-main.466.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Clement Rebuffel, Thomas Scialom, Laure Soulier, Benjamin Piwowarski, Sylvain Lamprier, Jacopo Staiano, Geoffrey Scoutheeten, and Patrick Gallinari. Data-QuestEval: A referenceless metric for data-to-text semantic evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 8029–8036, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.633.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Michael Schlichtkrull, Thomas Kipf, Peter Bloem, Rianne Berg, Ivan Titov, and Max Welling. *Modeling Relational Data with Graph Convolutional Networks*, pp. 593–607. 06 2018. ISBN 978-3-319-93416-7. doi: 10.1007/978-3-319-93417-4_38.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2015. doi: 10.1109/cvpr.2015.7298682. URL http://dx.doi.org/10.1109/CVPR.2015.7298682.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704.

Thibault Sellam, Amy Pu, Hyung Won Chung, Sebastian Gehrmann, Qijun Tan, Markus Freitag, Dipanjan Das, and Ankur Parikh. Learning to evaluate translation beyond English: BLEURT submissions to the WMT metrics 2020 shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pp. 921–927, Online, November 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.102.

Anastasia Shimorina, Claire Gardent, Shashi Narayan, and Laura Perez-Beltrachini. WebNLG Challenge: Human Evaluation Results. Technical report, Loria & Inria Grand Est, January 2018. URL https://hal.archives-ouvertes.fr/hal-03007072.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *CoRR*, abs/2104.06644, 2021. URL https://arxiv.org/abs/2104.06644.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pp. 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297, 2020. URL https://arxiv.org/abs/2004.09297.

Kristina Toutanova, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon. Representing text for joint embedding of text and knowledge bases. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1499–1509, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1174. URL https://aclanthology.org/D15-1174.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 2071–2080. JMLR.org, 2016.

Zhigang Wang and Juanzi Li. Text-enhanced representation learning for knowledge graph. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pp. 1293–1299. AAAI Press, 2016. ISBN 9781577357704.

Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2263, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1239. URL https://aclanthology.org/D17-1239.

Jiawei Wu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Knowledge representation via joint learning of sequential text and knowledge graphs. *arXiv preprint arXiv:1609.07075*, 2016.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. Cross-lingual knowledge graph alignment via graph matching neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3156–3161, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1304. URL https://aclanthology.org/P19-1304.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 250–259, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1025. URL https://aclanthology.org/K16-1025.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and li Deng. Embedding entities and relations for learning and inference in knowledge bases. 12 2014.

Haiyang Yu, Ningyu Zhang, Shumin Deng, Hongbin Ye, Wei Zhang, and Huajun Chen. Bridging text and knowledge with multi-prototype embedding for few-shot relational triple extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6399–6410, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.563. URL https://aclanthology.org/2020.coling-main.563.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Impact of Batch Size

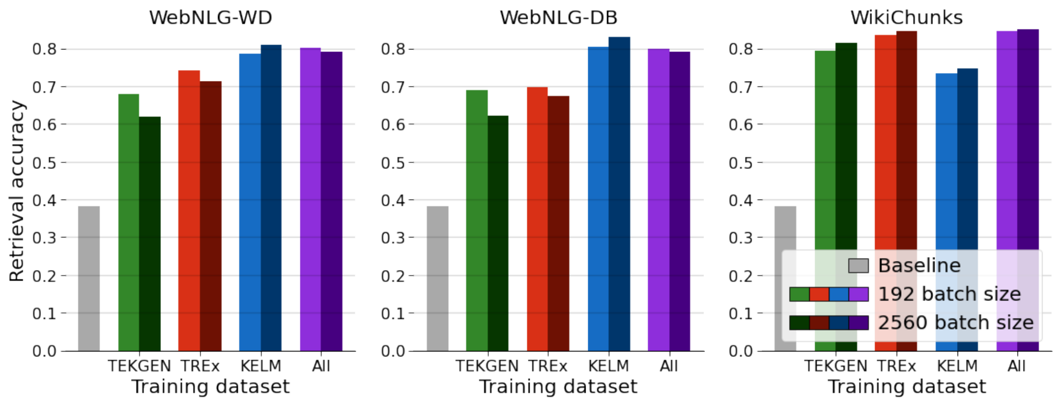

Figure 7: **Small vs. Large Batch Size.** Large batch sizes help a little on data with lower alignement quality (WikiChunks). Overall, the improvement is inconsistent.

B Scatter Plot Comparison of BLEURT and our Metric

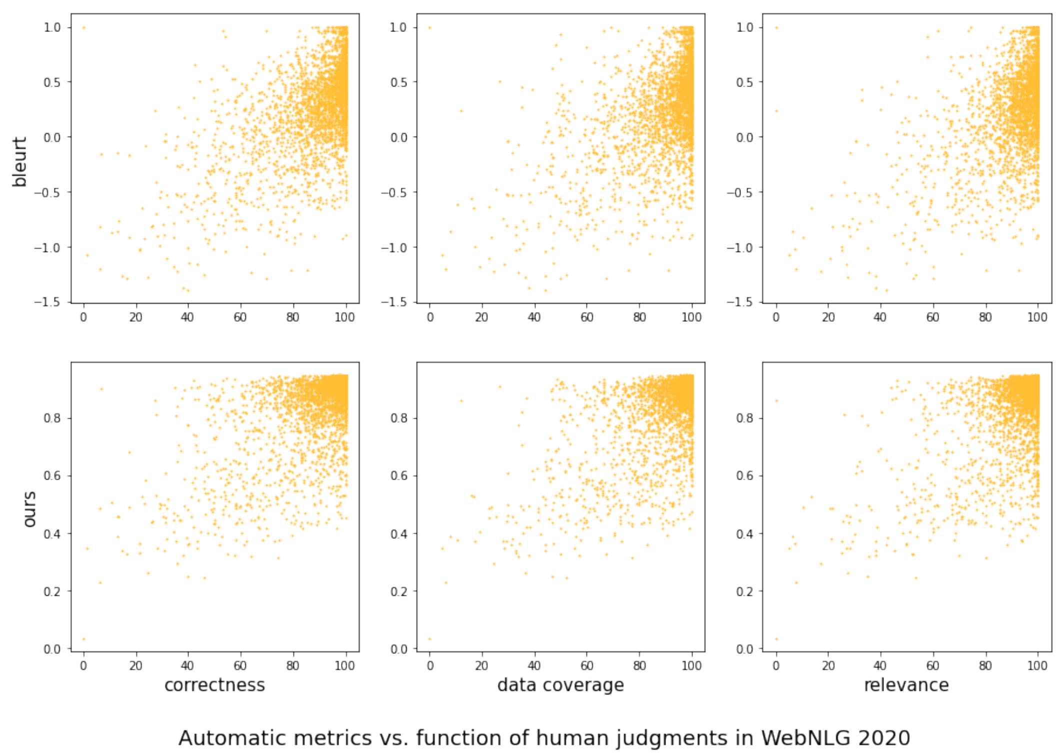

Figure 8: **Human judgment and automated evaluation values for every point in WebNLG 2020**. Contrary to BLEURT, our metric does not require a reference. Still, their correlations to human judgments are on par with each other, as our metric has better Pearson correlations and worse Spearman correlations.

C SYMMETRICAL RELATIONSHIPS IN WEBNLG

We manually inspected all relationships in WEBNLGand deemed the following to be symmetrical in nature:

"taxon synonym", "partner in business or sport", "opposite of", "partially coincident with", "physically interacts with", "partner", "relative", "related category", "connects with", "twinned administrative body", "different from", "said to be the same as", "sibling", "adjacent station", "shares border with"