DBpedia GSoC 2023 Project Report

Kancharla Aditya Hari

September 23, 2023

Abstract

RDF-to-Text generation is a subproblem of data-to-text generation and involves generating natural language text using RDF triples. This is an important part as RDF triples are commonly used to in knowledge graphs such as those maintained by Wikidata and DBpedia. These knowledge graphs represent an extensive web of knowledge, and through RDF-to-Text can be used to generative informative content. Knowledge graphs can be considered to be agnostic to their language of expression, and thus can be used to enrich content available in low and medium resource languages by leveraging resources available in high-resource languages. In this project, we focus on the problem of multilingual RDF-to-Text generation, in which the RDF triples in English are verbalized in different languages. This problems suffers from lack of availability of training data for low-resource languages. We address this problem by systematically denoising synthetically generated data in an iterative manner which shows significant improvements over using only the available datasets.

1 Introduction

Data-to-text is a problem in NLG in which natural language text is generated from structured data. RDF-to-Text (R2T) is subproblem of this in which structured RDF triples in the form of subject-predicate-object are used to generate natural language text. This can be useful for generating informative texts such as Wikipedia articles. This can be used to verbalize knowledge graphs, which use graphs to encode a wide variety of information. An important property of knowledge graphs is that they are language agnostic, with the encoded information being independent of the surface realization. This motivates the problem of multilingual RDF-to-Text (mR2T), in which RDF triples are used to generate text in multiple languages. This can be used to generate informative content for low-resource languages which typically have lower availability of such content by leveraging knowledge graphs such as Wikidata and DBpedia which are usually verbalized in English.
However, mR2T suffers from the problem of lack of availability of quality training data. Availability of training data for languages other than English is typically limited. The most widely used multlingual source of data is the WebNLG dataset, which includes data for English, Russian and several low-resource languages, namely Irish, Welsh, Maltese, and Breton.

Language	Train	Validation	Test
English	35426	1667	1779
Russian	14630	2065	1102
German	17943	868	868
Irish	35426	1667	1779
Hindi	50923	1344	421

Table 1: Dataset statistics. Hindi sources from XAlign, the rest are sourced from WebNLG

The Russian data was created by machine translating and post-editing the English dataset, while the low-resource data is an order of magnitude smaller in size (1500 vs 30000). Thus, in this project we explore solutions for mR2T based on using augmented datasets created by generating synthetic training pairs. We first propose a method for automatically generating RDF-sentence pairs which aligns RDF triples extracted from DBpedia with natural language text in different languages extracted from Wikipedia. This dataset in conjunction with the trusted datasets can be used to train models on greater amounts of data. However, due to the automatic nature of its construction, it is likely to contain noisy data that will hinder performance. Thus we also investigate a method which uses the data in a more careful way by quantifying the noise in each sample and iteratively training the model on cleaner data. We summarize our work and the results of the methods explored in this work. The generated synthetic dataset, the generated outputs, and the code to replicate the work are made publicly available.

2 Datasets Used

Two datasets are used for experiments - the WebNLG dataset and XAlign dataset.
The English, Russian, Irish, and German language variants of the WebNLG dataset are used. These contains human annotated RDF triple-sentence pairs for English and Irish, human post-edited machine-translated pairs for Russian, and machine-translated pairs for German. The RDF triples are sourced from DBpedia.
The Hindi data of the XAlign dataset is used, which contains synthetically generated training and validation data and human annotated testing data. Here, the RDF triples are sourced from Wikidata. Note that our approach requires a small, trusted dataset for training, which is not available in the XAlign dataset. Thus, we split the test data, with one half held-out for testing and the other used for the training process. We also experimented with using a machine translated version of the English WebNLG dataset as the trusted source instead.
The dataset statistics are summarized in 1

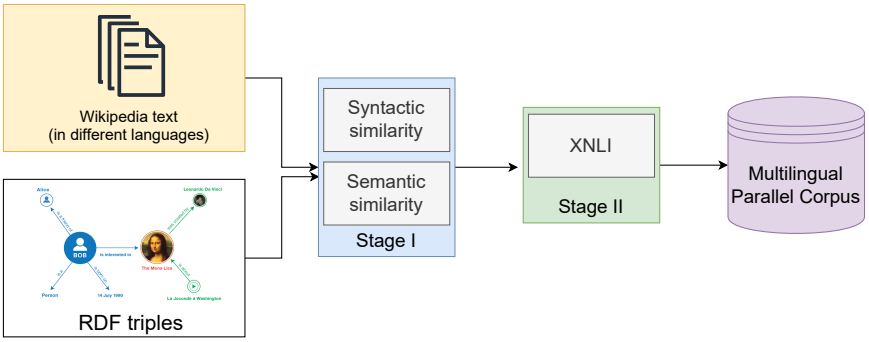


Figure 1: Two stage multilingual aligner

3 Synthetic Corpus

The first contribution of this project is a synthetic dataset of RDF triple-sentence pairs. For this, sentences from Wikipedia and RDF triples from DBpedia are automatically aligned, which is described in the following sections. In this section, we describe the creation of the synthetic corpus.

RDF Triples - The RDF triples that the sentences will be aligned with are extracted from the DBpedia Ontology. These are manually created, high quality properties. Here, the heads are identified using resource tags from only the English variant of DBpedia. The English labels of the entities and properties are used rather than the resource IDs

Article texts - The sentences which we will align with the RDF triples are extracted from Wikipedia. The Wikipedia Abstracts dataset is used for this, which contains the introduction section of Wikipedia articles. Note that this contains the DBpedia resource ID specific to the language. The article for the German footballer Toni Kroos, for example, in German and in English are the same for our purposes and will map to the same properties, and thus need to be unified.

Entities from the following ontology domains are extracted - Place, Building, Infrastructure, Person, Organization, and Work. An example of the extracted data is shown in 3

4 Automatic Alignment

The synthetic data is generated by automatically aligning RDF triples extracted from a knowledge base (DBpedia in this case) with natural language text (from Wikipedia in this case). The approach used in this work is based on the approach proposed by XAlign [1]. In this, a two-stage approach is used. In the first stage, candidates are generated by scoring sentences on their similarity to RDF triples and selecting the top-K within a threshold. In the second stage, candidates are selected by determining if the facts are entailed by the corresponding sentences. We describe the steps in detail next. The process is summarized in 1

4.1 Candidate Generation

As mentioned above, candidates are generated in this step by scoring the similarity of sentences with RDF triples. A combination of two similarity measures is used - syntactic and semantic similarity.

Syntactic similarity

Syntactic similarity refers to similarity based on the similarity of grammatical structure or syntax of two texts. This typically relies on features such as arrangement and frequency of words independent of their meaning. We use TF-IDF similarity between the linearized RDF triples and the sentences as the measure of syntactic similarity. Note that while our RDF triples are always expressed in English, the sentences can be in different languages. Thus, we translate the sentences using Meta's No Language Left Behind's model [2] and use the translated versions for computing TF-IDF similarity. The distilled, 600M parameter model available on Huggingface was used.

Semantic similarity

In contrast to syntactic similarity, semantic similarity attempts to score the similarity of two texts based on their meaning rather than syntax. This is necessary as sentences with similar structures can often have very different meanings, while sentences with different structure can have similar meaning. In this work, we use cosine similarity of the sentence embeddings as the measure of semantic similarity. Specifically, we use sentence embeddings generated through SentenceTransformer's distilled RoBERTa model. We experimented with a few methods of further refining these embeddings.

- Contrastive learning based finetuning (FT) - To align the sentence embeddings with the RDF embeddings, we train the model on the task of semantic textual similarity using a contrastive learning. We use the fact-sentence pairs available in the WebNLG dataset as the positive samples, and generate negative samples by randomly permuting these pairs. 37,000 positive examples and 500,000 synthetically generated negative samples are present in the dataset. This in theory decreases the distance between the positive pairs and decreases the distance between negative pairs.

- Domain adaption through continued pre-training (PT) - The model used is pretrained on only natural text. Thus, it might struggle with generating embeddings for RDF triples. Thus we perform domain adaption by continued pre-training of the model on the next token prediction task on a corpus of only RDF triples. We also experimented with pretraining models from scratch using custom tokenizers.

Language	RDF input	Sentence	English translation
English	Arsenal F.C.-manager-Mikel Arteta	On 20 December 2019, Arsenal appointed former club captain Mikel Arteta as the new head coach.	
Russian	Proaza | subdivision | Asturias <TSP> Proaza | country | Spain <TSP> Proaza | type | Municipalities of Spain	Проаса (исп.Proaza) — муниципалитет в Испании, входит в провинцию Астурия.	Proaza (Spanish: Proaza) is a municipality in Spain, part of the province of Asturias.
German	Bill Stevenson (musician) | givenName | John William Stevenson <TSP> Bill Stevenson (musician) | birthPlace | Torrance, California	John William „Bill" Stevenson (* 10. September 1963 in Torrance, Kalifornien) ist ein US-amerikanischer Musiker und Musikproduzent.	John William "Bill" Stevenson (born September 10, 1963 in Torrance, California) is an American musician and record producer
Irish	Matthias Bachinger | birthDate | 1987-04-02	Rugadh é ar an 2 Aibreán 1987.	He was born on 2 April 1987
Hindi	Nayan Mongia | occupation | cricketer <TSP> Nayan Mongia | countryOfCitizenship | India	वे पूर्व भारतीय क्रिकेटर हैं।	He is a former Indian cricketer.

Figure 2: Examples of automatically aligned data in different languages

Language	Entities	Sentences per entity	Facts per sentence
English	408306	1.55	1.46
Russian	362648	1.35	1.50
German	262774	1.29	1.51
Irish	18741	1.31	2.06

Table 2: Statistics of synthetic dataset

Results

To measure the performance of these similarity metrics, we use the same semantic textual similarity task described above. Using only syntactic similarity results in an F1 score of 0.82. To combine the syntactic and semantic scores, a kNN classifier is used with $k = 3$. The F1 scores of the different semantic methods is listed in table 3. A combination of the syntactic score and the semantic score computed using the sentence transformer model fine-tuned on the STS task reports the best performance and is used for candidate generation. The statistics of the resultant dataset after candidate generation is reported in table 2

4.2 Candidate Selection

In this stage, the candidate RDF triple-sentence pairs are filtered to select the final set of pairs that will serve as the synthetic dataset. For this, transfer learning using the natural language inference (NLI) task is used. In NLI, given a premise and a hypothesis, the task is to classify the hypothesis as an entailment, contradiction, or neutral to the premise. The alignment of RDF triples to sentences is similar to this task, with the RDF triple forming the hypothesis and the sentence the premise. For a given sentence, all the subset of RDF

Method	Base	Base + Syntactic
RoBERTa base	0.74	0.82
SentenceTransformer (ST)	0.78	0.83
ST + PT	0.82	0.84
ST + FT	0.87	0.89
ST + FT + PT	0.87	0.87

Table 3: Similarity scores results

triples which are classified as entailments are taken to form the final pair. We used an mT5-large model finetuned on the XNLI task available on Huggingface for this step. This allows this process to be done cross-lingually, with the RDF triple expressed in English and the sentence in different languages. The statistics of the final synthetic dataset after candidate selection is reported in table 3. Some examples are shown in figure 2

5 Denoising Noisy Data

In machine learning, noise refers to irrelevant or meaningless information in the training data that result in models generalizing to incorrect patterns or hindering their ability to identify patterns. In our dataset, this can be in the form of candidates where the alignment is completely incorrect or where the sentence or the RDF triple contain insufficient or extra information. These can result in poor performance without careful training. For instance, in the English example in 2, while the fact that, it also contains other information such as "former club captain" and the date of appointment. However, it still contains useful information that can be leveraged by the model to identify how the RDF triple can be verbalized.

This motivates our iterative training procedure. The problem of of learning from noisy data has been explored in literature, and systematic use of the noisy data can be used to improve the performance of models while mitigating the impact of noise. To this end, we first quantify the noise in any given pair of RDF triple and sentence. Then, an iterative training approach which progressively trains on the model on the higher quality examples is used. We describe these steps in detail in the following section.

5.1 Quantifying Noise

We use the denoising approach proposed in [3]. Assume that we have a model M_θ parameterized by θ that given a RDF triple x and sentence y outputs the probability $p(y|x, \theta)$ that the sentence corresponds to the RDF triple. Then, noisy log probability of the pair (x, y) can be computed as

$$L_{p(y|x,\theta)} = log(p(y|x, \theta))$$

Consider two models - a noisy model M_θ, and a denoised model $M_{\hat{\theta}}$ which outputs a more accurate probability distribution. Then, the quality of a given pair (x, y) can be quantified as

$$u = quality(x, y, \theta, \hat{\theta}) = L_{p(y|x,\theta)} - L_{p(y|x,\hat{\theta})}$$

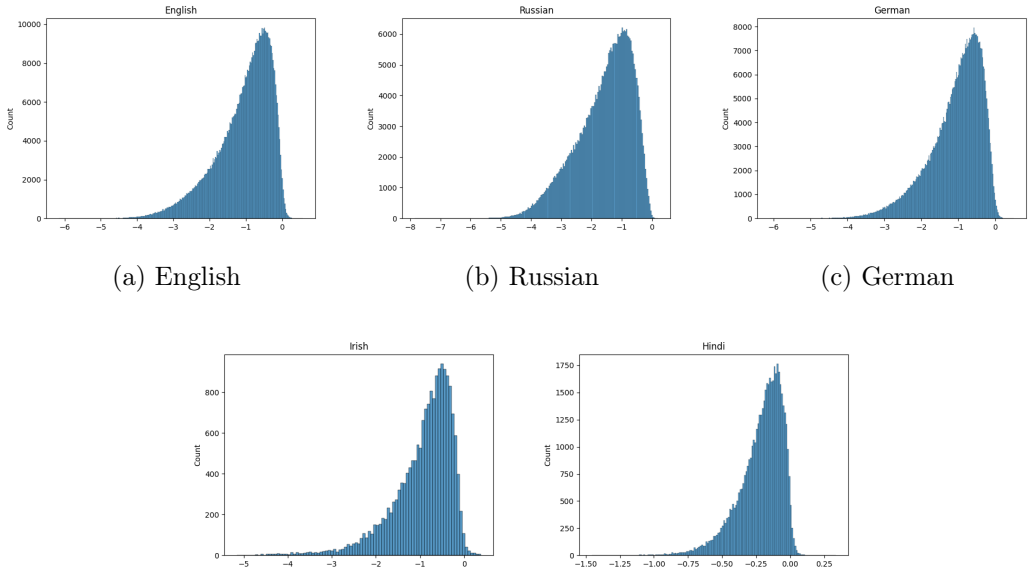
A positive score means that the pair is more likely according to the denoised model than the noisy model, indicating the it is of higher quality. To estimate $p(y|x, \theta)$, we use seq2seq language models, specifically the t5-small model. A model is first trained on the noisy data $\hat{D}$ to obtain the noisy model. This model is then further trained on the smaller, trusted

Language	Mean	Std	>0 percentage
English	1.658	0.622	0.99994
Russian	1.794	0.688	0.99992
German	1.783	0.665	0.99989
Irish	1.871	0.776	0.99994

Table 4: Difference of score from noisy model and denoised model. A positive score indicates higher score by denoised model

dataset D to obtain the denoised model.

5.1.1 Results



(a) English (b) Russian (c) German

(d) Irish (e) Hindi

Figure 3: Distribution of scores in synthetic dataset.

The distribution of the quality of the samples in the generated synthetic datasets is shown in figure 3. Only a few samples have a score above 0. Additionally, for a soundness check we also compute the quality of each sample in the trusted data, the results of which are reported in 4. The last column represents the percentage of samples that have a score greater than zero, which indicate a high quality example. This shows that the trusted data contains almost exclusively high quality examples, which is as expected.

6 Iterative Training

The iterative training approach is based on the key assumption that exposing the model to progressively higher quality data will allow to effectively learn from the noisy data without

hindering performance. To accomplish this, an iterative training procedure in which the model is trained in different stages, with each stage involving training on progressively higher quality data. Specifically, the model is first trained on data sampled from the top 50%, then on data sampled from the top 25%, and finally on only the trusted data. Note that the trusted data is also included in the first two stages. Model selection is done based on validation BLEU score. A pretrained t5-small model is used for English, and mt5-small for the other languages. A linearly decaying learning rate of 3e-5 is used for each stage. For a baseline, the model is trained on the trusted data.

6.1 Ablations

To validate the approach, two ablations were performed which focus on understanding how the models perform without the iterative training procedure.

Non-iterative training on higher quality data - In this, the models were trained directly on the higher quality noisy data without initially training it on the lower quality data.

Training on only noisy data - The trusted data is included in each step of the training process, as per previous denoising approaches. To study the benefits of the noisy data, we also train the models in an iterative manner on only the noisy data before training on the trusted data in the final step.

Model	BLEU	METEOR	CHRF
Trusted data	44.284	60.638	65.387
Top half	44.725	60.714	65.651
Top quarter	45.079	61.164	65.965
Only trusted	45.04	60.59	65.569

Table 5: Results after iterative training for English

Model	Type	BLEU	METEOR	CHRF
Baseline	1	53.894	64.194	69.703
	2	44.6	61.756	66.845
	3	35.371	58.202	61.015
Top half	1	55.687	65.094	70.991
	2	41.925	61.437	65.808
	3	36.006	58.002	60.957
Top quarter	1	55.648	65.375	71.045
	2	44.257	62.84	66.462
	3	35.454	58.126	61.361
Only trusted	1	56.247	65.473	70.976
	2	42.972	61.734	65.665
	3	35.995	57.417	60.855

Table 6: Results by input type for English

6.2 Results

6.2.1 Iterative training

WebNLG - English, Russian, German, Irish

The results of the iterative training process for all the languages in the WebNLG dataset - English and Russian, German and Irish, are listed in 5 and 7 respectively. In each case, the best performance is reported by the model trained iteratively on the denoised synthetic data. However, the performance gain between stages in minimal in most cases and in the case of German, the best performance is reported by the model trained on the 75th percentile data. This indicates that while the synthetic data results in improved performance, the iterative process's contribution requires more analysis.

For English, the performance is also reported by the input type - seen entities (type 1),

Model	BLEU			METEOR			CHRF		
	Russian	German	Irish	Russian	German	Irish	Russian	German	Irish
Baseline	50.705	57.256	5.471	59.421	67.191	15.582	69.298	72.525	25.474
Top half	48.579	56.593	5.871	56.599	66.477	17.438	66.056	70.983	26.369
Top quarter	50.025	58.748	6.231	57.988	68.051	16.699	67.386	72.679	26.332
Only trusted	52.359	58.451	6.593	60.309	68.542	18.463	69.874	72.85	26.491

Table 7: Results after iterative training for Russian, German, and Irish

Model	BLEU	METEOR	CHRF
Baseline	29.542	53.245	55.502
Bottom half	10.972	38.651	43.423
Top half	21.614	42.376	45.364
Top quarter	41.808	54.061	59.669

Table 8: Results after iterative training for Hindi

unseen entities from seen categories (type 2), and unseen entities from unseen categories (type 3) in table 6. Compared to the baseline, the models trained on the synthetic data report higher performance on the unseen categories, indicating their greater ability to generalize to unseen data. However, their performance on unseen entities from seen categories is significantly worse.

XAlign - Hindi

The results for the iterative training process for the Hindi data from the XAlign dataset is reported in 8. Here, the merits of the iterative process are more significant than in the WebNLG dataset. There is sizeable improvement in performance after each stage, and the final stage results in significantly better performance than the baseline.

Using the translated version of the English WebNLG results in significantly worse peformance, with a BLEU score of only 10 at the end of the training process. Thus, the numbers reported here are with a partition of the test split used as the trusted dataset.

6.2.2 Ablations

English - The results for the ablation studies on English are reported in table 9 and table 10. In each case, the models were trained in a non-iterative manner i.e trained on each data split separately. Table 9 reveals that training on the cleaner data does indeed result in better performance. However, the difference is insignificant in table 10. Remember that the difference between the two is that in the former, only noisy data is included in each split, whereas the latter also includes the trusted data. This indicates including the trusted data in the training stages results in a saturation of performance, with the improvement in performance largely a consequence of the additional data rather than the iterative process.

Hindi - Due to the absence of trusted data in Hindi, only the 2nd ablation is performed for Hindi. The results are reported in table 11. Again, training on the clear data results in improved performance.

Model	BLEU	METEOR	ROUGE1	CHRF
Top half	6.92	21.618	40.894	30.004
Top quarter	7.755	21.434	41.497	31.979

Table 9: Results after training on only noisy data non-iteratively for English

Model	BLEU	METEOR	ROUGE1	CHRF
Top half	44.725	60.714	50.348	65.651
Top quarter	44.761	61.001	50.309	65.573

Table 10: Results after non-iterative training for English

Model	BLEU	METEOR	CHRF
Bottom half	10.972	38.651	43.423
Top half	21.614	42.376	45.364
Top quarter	31.724	45.046	50.566

Table 11: Results after non-iterative training for Hindi

7 Input RDF organization

Further two experiments were performed to study the role of organizing the input.

- Input expression - The input to our models can be thought of as a linearized knowledge graph. Thus, the form in which this knowledge is expressed can vary. Two formulations were experimented with - in one, the RDF triples are verbalized as "subject | predicate | object", with multiple triples being concatenated using a special <TSP> token. In the other, the input was expressed as "<H> subject <R> predicate <T> object". No special token to separate triples is required in this formulation
- Input RDF ordering - The synthetic dataset contains concatenated RDF triples. The order of concatenation is arbitrary. So, an experiment was performed to understand if this arbitrary ordering has an impact on the performance. For this, two models were trained - one with the trusted dataset with the triples in the given order, and one with the trusted dataset with the order of the triples randomized.

7.1 Results

For input expression, the H-R-T expression resulted in a BLEU score of 42, whereas the TSP expression resulted in a BLEU score of 44. Thus, the further experiments the results of which are reported above were conducted using the TSP expression.
For RDF ordering, both experiments resulted in a BLEU score of 42. Thus, input ordering was not experimented with for the further experiments.

8 Domain Adaptive Translate-Test

Motivated by recent literature which suggests that the translate-test approach, in which multilingual generation is achieved by simply translating the outputs of a single-language model, a simple translate-test based approach was also investigated as part of this project. In this, the NLLB machine translation model was finetuned on sentence pairs in English

Language	BLEU	chrf++	TER
Maltese	16.49	0.47	0.7
Welsh	20.97	0.49	0.67
Irish	15.66	0.44	0.73
Russian	36.01	0.57	0.67

Table 12: Results for WebNLG 2023 challenge using domain adaptive translate-test

and the different low resource languages, to accomplish the means of domain adaption. This was then used to translate the outputs of a T5-small model finetuned on only the English data to generate multilingual data in the required languages.
The results of this approach were submitted to the WebNLG 2023 workshop, and can be found in table 12. Note the metrics used here those used in the WebNLG 2023 challenge, and thus different from the metrics reported for the previous experiments. Irish here reports significantly better performance than generating directly in Irish as in previous experiments, while Russian results in significantly worse performance. This points to the importance of availability of high quality data, with native generation in Irish suffering.

9 Conclusion

Multilingual generation of natural language is an important problem in the field on NLP, and will only gain significance as we move toward a more digitally accessible world for everyone. Improving the quality of content available for languages other than English, particularly low-resource languages, is paramount. In this project we experiment with a wide array of methods to study this problem, and presented a way to automatically generate training data for multiple languages by leveraging existing resources in English. We also presented a new training paradigm that shows promising results using the synthetic data.
The training code, generated synthetic data, and generated outputs have been made publicly available on GitHub

10 Limitations and Future Work

- More comprehensive evaluation - While we have performed thorough evaluation of our models in multiple languages and with multiple ablations, there is still room for evaluation. For instance, studying the performance based on the relation types that an input features, based on the length of the reference are two ways in which the quality of the synthetic data can be analyzed. Additionally, metrics such as PARENT score, which evaluates generations based on not just the reference but also the input, can also be used.

- Multisentence synthetic dataset - The synthetic data generated as part of this project

only features single sentence samples. For generating longer content, it's important to have training data of similar quality. Thus, generating synthetic data of longer lengths is a potential avenue for future explorations.

- Synthetic data coreference resolution - The examples in figure 2 show that in many cases, the coreferenences are not resolves e.g "He is a former Indian cricketer" rather than "Nayan Mongia is a former Indian cricketer". This results in divergence from the trusted data, and poorer quality data. Performing coreference resolution prior to scoring and aligning the data can improve the quality of the synthetic data.

- Multilingual models - The experiments performed here train separate models for every language. This results in a necessitiy of availability of training data in each language. Multilingual generation requires careful analysis of the included languages, as training with very different languages can result in worse performance for each language. Due to time constraints, it was not feasible to study this as part of this project. However, in the future this dimension can be added, which will have the benefit of not requiring trusted data for each language.

References

[1] Tushar Abhishek et al. "XAlign: Cross-lingual Fact-to-Text Alignment and Generation for Low-Resource Languages". In: *CoRR* abs/2202.00291 (2022). arXiv: 2202.00291. URL: https://arxiv.org/abs/2202.00291.

[2] NLLB Team et al. *No Language Left Behind: Scaling Human-Centered Machine Translation*. 2022. arXiv: 2207.04672 [cs.CL].

[3] Wei Wang et al. "Denoising Neural Machine Translation Training with Trusted Data and Online Data Selection". In: *CoRR* abs/1809.00068 (2018). arXiv: 1809.00068. URL: http://arxiv.org/abs/1809.00068.