Robustification of Multilingual Language Models to Real-world Noise in Crosslingual Zero-shot Settings with Robust Contrastive Pretraining

Asa Cooper Stickland[⋆1,2] Sailik Sengupta[⋆1]

Jason Krone[1] He He[1,3] Saab Mansour[1]

[1]aWS AI Labs

[2]University of Edinburgh

[3]New York University

EACL 2023

🇭🇷 Dubrovnik

Text Classification

Sentence-level classification task (eg. Intent Classification, XNLI, etc.)

- Croatia is such a lovely place → +ve

Token-level classification task (eg. NER, Slot-labeling):

- **Croatia** is such a lovely place → {**Croatia**: Country}

Text Classification

Sentence-level classification task (eg. Intent Classification, XNLI, etc.)

- Croatia is such a lovely place → +ve
- Croatia is suhc a lovely place → ?

Token-level classification task (eg. NER, Slot-labeling):

- **Croatia** is such a lovely place → {**Croatia**: Country}
- **Croatia** is suhc a lovely place → ?

🤔 What happens when faced with real-world noise?

In this work, we study this question for languages beyond English.

Table of Contents

1. Related Work
2. Evaluation Mechanisms
   - Finding Noisy Data
   - Creating Evaluation Test-sets
   - Multilingual Noise Characteristics
3. Robust Constrastive Pretraining
4. Experiments
   - Experimental Setup
   - Robustness of Multilingual Models
   - A Study of Errors

Related Work

- *Works have investigated the impact of various noise types, mostly for English* – misspellings [BB17, KLEG19, MKS21], casing [vMvdLCFK20], paraphrases [EGMS19], morphological variance [TJKS20], synonyms [SKM21], dialectical variance [SLS+22]
- *Methods to improve robustness of SOTA models have considered* – Data augmentation during pre-training [TJKS20, SLS+22] or the task-training stage [PLZ+21], token-free models motivate robustness in multilingual settings [CGTW21, XBC+22, TTR+21], Adversarial Logit Pairing [EGMS19]
- *Our works is similar to works in computer vision that have used of Contrastive learning to boost model robustness* [FLC+21, GL21, JCCW20, KTH20]

Table of Contents

1. Related Work

2. Evaluation Mechanisms
   - Finding Noisy Data
   - Creating Evaluation Test-sets
   - Multilingual Noise Characteristics

3. Robust Constrastive Pretraining

4. Experiments
   - Experimental Setup
   - Robustness of Multilingual Models
   - A Study of Errors

Finding Noisy Data

There is a lack of benchmark to investigate the robustness of multilingual models. *Why?* 🤔

Finding Noisy Data

There is a lack of benchmark to investigate the robustness of multilingual models. *Why?* 🤔

Synthetic noise-generation methods (mostly developed for English) need linguistic expertise to create benchmarks for individual languages.
What to do? 🤔

Finding Noisy Data

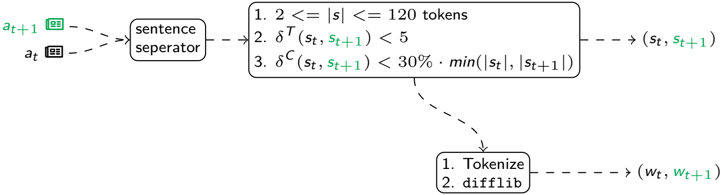
There is a lack of benchmark to investigate the robustness of multilingual models. *Why?* 🤔

Synthetic noise-generation methods (mostly developed for English) need linguistic expertise to create benchmarks for individual languages.
What to do? 🤔

Can we find a data source from where we can obtain such data?

💡 Wikipedia articles are continually updated/edited. Maybe we can mine these edits. (We also leverage other corpora such as Lang8.)

Finding Noisy Data



a_{t+1}

a_t

sentence seperator

1. $2 <= |s| <= 120$ tokens
2. $\delta^T(s_t, s_{t+1}) < 5$
3. $\delta^C(s_t, s_{t+1}) < 30\% \cdot min(|s_t|, |s_{t+1}|)$

(s_t, s_{t+1})

1. Tokenize
2. `difflib`

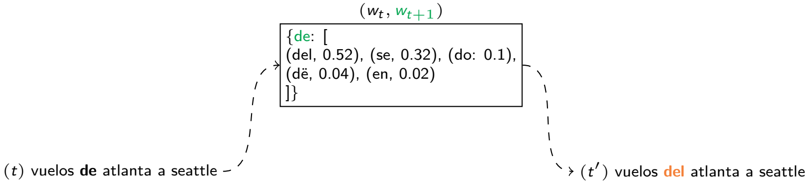
(w_t, w_{t+1})

Similar to [TMKK20], we obtain sentence edit dictionaries and word-edit dictionaries.

Creating Evaluation Test-sets

💡 Use word-edit dictionaries for noisy test-set creation!

We note that this makes our test-data limited to work level edits. But, we can have multiple words manipulated in a single utterance.



(w_t, w_{t+1})

{de: [
(del, 0.52), (se, 0.32), (do: 0.1),
(dë, 0.04), (en, 0.02)
]}

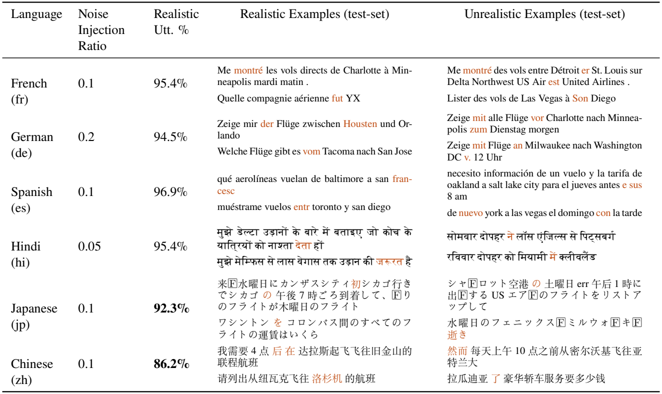
(t) vuelos **de** atlanta a seattle

(t') vuelos **del** atlanta a seattle

Creating Evaluation Test-sets – QA

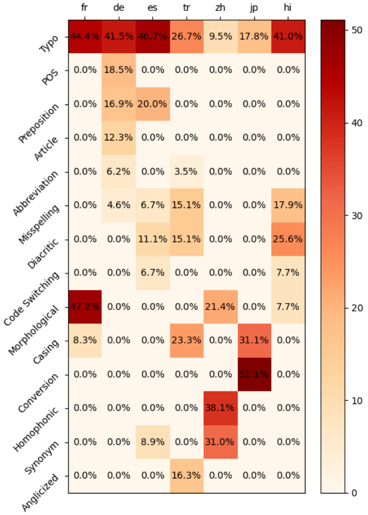
We inject various degrees of noise and conduct evaluation to decide which test sets are more realistic.

We keep test-data on if they have $< 5\%$ unrealistic errors.



Language	Noise Injection Ratio	Realistic Utt. %	Realistic Examples (test-set)	Unrealistic Examples (test-set)
French (fr)	0.1	95.4%	Me montré les vols directs de Charlotte à Minneapolis mardi matin . Quelle compagnie aérienne fut YX	Me montré des vols entre Détroit er St. Louis sur Delta Northwest US Air est United Airlines . Lister des vols de Las Vegas à Son Diego
German (de)	0.2	94.5%	Zeige mir der Flüge zwischen Housten und Orlando Welche Flüge gibt es vom Tacoma nach San Jose	Zeige mit alle Flüge vor Charlotte nach Minneapolis zum Dienstag morgen Zeige mit Flüge an Milwaukee nach Washington DC v. 12 Uhr
Spanish (es)	0.1	96.9%	qué aerolíneas vuelan de baltimore a san francesc muéstrame vuelos entr toronto y san diego	necesito información de un vuelo y la tarifa de oakland a salt lake city para el jueves antes e sus 8 am de nuevo york a las vegas el domingo con la tarde
Hindi (hi)	0.05	95.4%	मुझे डेल्टा उड़ानों के बारे में बताइए जो कोच के यात्रियों को नाश्ता देता हों मुझे मेम्फिस से लास वेगास तक उड़ान की जरूरत है	सोमवार दोपहर ने लॉस एंजिल्स से पिट्सबर्ग रविवार दोपहर को मियामी में क्लीवलैंड
Japanese (jp)	0.1	**92.3%**	来🄵水曜日にカンザスシティ初シカゴ行きでシカゴ の 午後 7 時ごろ到着して、🄵りのフライトが木曜日のフライト ワシントン を コロンバス間のすべてのフライトの運賃はいくら	シャ🄵ロット空港 の 土曜日 err 午后 1 時に出🄵する US エア🄵のフライトをリストアップして 水曜日のフェニックス🄵ミルウォ🄵キ🄵逝き
Chinese (zh)	0.1	**86.2%**	我需要 4 点 后 在 达拉斯起飞飞往旧金山的联程航班 请列出从纽瓦克飞往 洛杉机 的航班	然而 每天上午 10 点之前从密尔沃基飞往亚特兰大 拉瓜迪亚 了 豪华轿车服务要多少钱

Multilingual Noise Characteristics



	fr	de	es	tr	zh	jp	hi
Typo	44.4%	41.5%	46.7%	26.7%	9.5%	17.8%	41.0%
POS	0.0%	18.5%	0.0%	0.0%	0.0%	0.0%	0.0%
Preposition	0.0%	16.9%	20.0%	0.0%	0.0%	0.0%	0.0%
Article	0.0%	12.3%	0.0%	0.0%	0.0%	0.0%	0.0%
Abbreviation	0.0%	6.2%	0.0%	3.5%	0.0%	0.0%	0.0%
Misspelling	0.0%	4.6%	6.7%	15.1%	0.0%	0.0%	17.9%
Diacritic	0.0%	0.0%	11.1%	15.1%	0.0%	0.0%	25.6%
Code Switching	0.0%	0.0%	6.7%	0.0%	0.0%	0.0%	7.7%
Morphological	47.2%	0.0%	0.0%	0.0%	21.4%	0.0%	7.7%
Casing	8.3%	0.0%	0.0%	23.3%	0.0%	31.1%	0.0%
Conversion	0.0%	0.0%	0.0%	0.0%	0.0%	51.1%	0.0%
Homophonic	0.0%	0.0%	0.0%	0.0%	38.1%	0.0%	0.0%
Synonym	0.0%	0.0%	8.9%	0.0%	31.0%	0.0%	0.0%
Anglicized	0.0%	0.0%	0.0%	16.3%	0.0%	0.0%	0.0%

Human evaluation of injected noise surfaces many interesting insights.

- Certain noise types are language specific (eg. `jp` has conversion, `tr` has anglicization errors).
- Certain noise types are common across languages (although `zh` has less typos due to `pinyin` style keyboards).

See our paper for more [Sec 3.3].

Table of Contents

1. Related Work

2. Evaluation Mechanisms
   - Finding Noisy Data
   - Creating Evaluation Test-sets
   - Multilingual Noise Characteristics

3. Robust Constrastive Pretraining

4. Experiments
   - Experimental Setup
   - Robustness of Multilingual Models
   - A Study of Errors

Robust Constrastive Pretraining

💡 Use sentence-edit dictionaries to pre-train multilingual models!

The intuition is that this will teach these multilingual models to represent incorrect and edited sentence closer to one another.

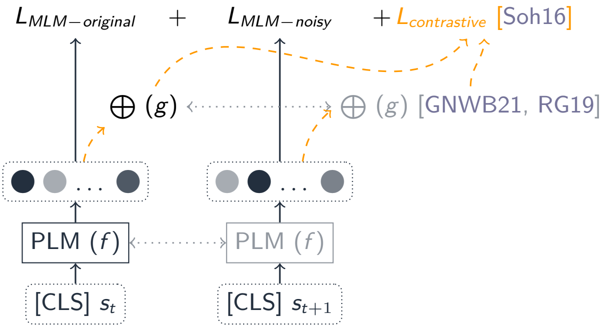

$L_{MLM-original}$ + $L_{MLM-noisy}$ + $L_{contrastive}$ [Soh16]
$\oplus$ (g)
$\oplus$ (g) [GNWB21, RG19]
...
...
PLM (f)
PLM (f)
[CLS] s_t
[CLS] s_{t+1}

Table of Contents

1. Related Work
2. Evaluation Mechanisms
   - Finding Noisy Data
   - Creating Evaluation Test-sets
   - Multilingual Noise Characteristics
3. Robust Constrastive Pretraining
4. Experiments
   - Experimental Setup
   - Robustness of Multilingual Models
   - A Study of Errors

Setup – Tasks, 0-shot Cross-lingual Transfer, Base models

The training data is the original english training set of each task.

Test data had two-splits for each lanaguage– the original test-set (Original) and the noise-added test-set (Noisy).

Dataset	Task	Training size (only en)	Languages (test)
MultiATIS++ [XHM20]	IC/SL	5k	de,en,es,fr,hi
+ training data aug. (en)		18k	de,en,es,fr,hi
MultiSNIPS	IC/SL	13k	en,es,fr,hi
+ training data aug. (en)		72k	en,es,fr,hi
WikiANN [PZM+17]	NER	20k	de,en,es,fr,hi,tr
XNLI [CRL+18]	NLI	392k	de,es,fr,hi,tr

Multilingual Model Robustness (as-is)

XLM-R$_{\textbf{base}}$ [CKG+20] > m-BERT [DCLT19] > Canine-c [CGTW21]

Robustness of Multilingual Models

Task	Metric	XLMR	XLMR	XLMR	XLMR	XLMR	Gain
			+p(aug)	+t(En-aug)	+RCP	+RCP+t	
					(Ours)	(Ours)	
MultiATIS++	IC%	89.65	93.10	91.26	93.80	**94.57**	+4.92
	SL-F1	62.30	67.47	74.62	67.45	**80.68**	+18.38
MultiSNIPS	IC%	90.46	93.98	91.60	93.79	**94.53**	+4.07
	SL-F1	61.63	66.67	66.44	67.69	**70.20**	+8.57
Wiki-ann	NER-F1	69.48	72.32	-	**72.37**	-	+2.89
XNLI	NLI%	74.38	74.83	-	**75.06**	-	+0.68

RCP ↑ model robustness across all tasks metrics – Accuracy of IC & XNLI, F1-score for SL & NER (avg across languages).

Gains ↑↑ when agg. English noise data [SKM21] is used during task-time augmentation.

Robustness of Multilingual Models

Task	Metric	XLMR	XLMR +p(aug)	XLMR +t(En-aug)	XLMR +RCP (Ours)	XLMR +RCP+t (Ours)	Gain
MultiATIS++	IC%	89.65	93.10	91.26	93.80	**94.57**	+4.92
	SL-F1	62.30	67.47	74.62	67.45	**80.68**	+18.38
MultiSNIPS	IC%	90.46	93.98	91.60	93.79	**94.53**	+4.07
	SL-F1	61.63	66.67	66.44	67.69	**70.20**	+8.57
Wiki-ann	NER-F1	69.48	72.32	-	**72.37**	-	+2.89
XNLI	NLI%	74.38	74.83	-	**75.06**	-	+0.68

RCP ↑ model robustness across all tasks metrics – Accuracy of IC & XNLI, F1-score for SL & NER (avg across languages).

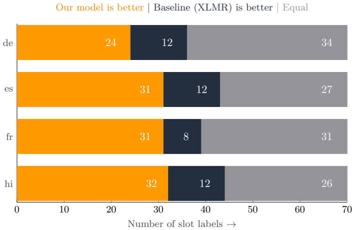
Gains ↑↑ when agg. English noise data [SKM21] is used during task-time augmentation.

RCP ↑ model performance on clean data too!

Task	Metric	XLMR	Ours	Gain
MultiATIS++	IC%	90.68	95.32	+4.64
	SL-F1	71.45	84.07	+12.62
MultiSNIPS	IC%	92.93	95.66	+2.73
	SL-F1	68.01	74.39	+6.38
Wiki-ann	NER-F1	74.14	76.34	+2.2
XNLI	NLI%	76.69	76.75	+0.06

A Study of Errors (on MultiATIS++)

Improvement in slot-label classification ($2\times$ `de`, $2.6\times$ `es`, `hi`, $4\times$ `fr`)



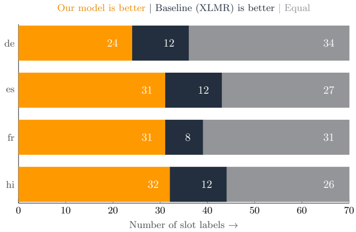
↑ Explicability of errors [OSK20]

👎 fromloc.airport_code → date

👍 fromloc.airport_code → toloc.airport_code

A Study of Errors (on MultiATIS++)

Improvement in slot-label classification (2× `de`, 2.6× `es`, `hi`, 4× `fr`)



↑ Explicability of errors [OSK20]

👎 fromloc.airport_code → date
👍 fromloc.airport_code → toloc.airport_code

We see a sharp drop in *hallucination* errors across all languages.

N/O	Model	de	es	fr	hi
Noisy	XLMR	315	358	413	671
	XLMR+RCP+t	21	123	33	204
Original	XLMR	208	262	334	460
	XLMR+RCP+t	19	106	22	180

↓ Hallucination errors

Model identifies irrelevant tokens as slot values. Eg.

"Ichs brauche einen Flug von Memphis nach Tacoma, der uber Los Angeles fliegt."

👍 O (über) ⟶ 👎 airline_code (uber)

Conclusion

- Multilingual test data to evaluate the robustness of multilingual models to noise.
- Performance of existing multilingual language models deteriorates on four tasks when tested on the noisy test data.
- Robust Contrastive Pretraining (RCP) can boost the robustness of existing multilingual language models.

Data & Code

https://github.com/amazon-science/multilingual-robust-contrastive-pretraining

Conclusion

- Multilingual test data to evaluate the robustness of multilingual models to noise.
- Performance of existing multilingual language models deteriorates on four tasks when tested on the noisy test data.
- Robust Contrastive Pretraining (RCP) can boost the robustness of existing multilingual language models.

Data & Code

https://github.com/amazon-science/multilingual-robust-contrastive-pretraining

✋ Questions? ✋

- Yonatan Belinkov and Yonatan Bisk, *Synthetic and natural noise both break neural machine translation*, arXiv preprint arXiv:1711.02173 (2017).
- Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting, *CANINE: pre-training an efficient tokenization-free encoder for language representation*, CoRR **abs/2103.06874** (2021).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, *Unsupervised cross-lingual representation learning at scale*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 8440–8451.

- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov, *XNLI: Evaluating cross-lingual sentence representations*, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (Brussels, Belgium), Association for Computational Linguistics, October-November 2018, pp. 2475–2485.

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, 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) (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019, pp. 4171–4186.

- Arash Einolghozati, Sonal Gupta, Mrinal Mohit, and Rushin Shah, *Improving robustness of task oriented dialog systems*, arXiv preprint arXiv:1911.05153 (2019).

- Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan, *When does contrastive learning preserve adversarial robustness from pretraining to finetuning?*, Advances in Neural Information Processing Systems **34** (2021).

- Aritra Ghosh and Andrew Lan, *Contrastive learning improves model robustness under label noise*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2703–2708.

- John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader, *DeCLUTR: Deep contrastive learning for unsupervised textual representations*, 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) (Online), Association for Computational Linguistics, August 2021, pp. 879–895.

- Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang, *Robust pre-training by adversarial contrastive learning*, Advances in Neural Information Processing Systems **33** (2020), 16199–16210.

- Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad, *Training on synthetic noise improves robustness to natural noise in machine translation*, arXiv preprint arXiv:1902.01509 (2019).

- Minseon Kim, Jihoon Tack, and Sung Ju Hwang, *Adversarial self-supervised contrastive learning*, Advances in Neural Information Processing Systems **33** (2020), 2983–2994.

- Pooya Moradi, Nishant Kambhatla, and Anoop Sarkar, *Measuring and improving faithfulness of attention in neural machine translation*, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (Online), Association for Computational Linguistics, April 2021, pp. 2791–2802.

- Alberto Olmo, Sailik Sengupta, and Subbarao Kambhampati, *Not all failure modes are created equal: Training deep neural networks for explicable (mis) classification*, ICML Workshop on Uncertainty and Robustness in Deep Learning (2020).

- Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao, *RADDLE: An evaluation benchmark and analysis platform for robust task-oriented dialog systems*, 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) (Online), Association for Computational Linguistics, August 2021, pp. 4418–4429.

- Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji, *Cross-lingual name tagging and linking for 282 languages*, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Vancouver, Canada), Association for Computational Linguistics, July 2017, pp. 1946–1958.

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

- Sailik Sengupta, Jason Krone, and Saab Mansour, *On the robustness of intent classification and slot labeling in goal-oriented dialog systems to real-world noise*, Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI (Online), Association for Computational Linguistics, November 2021, pp. 68–79.

- Soumajyoti Sarkar, Kaixiang Lin, Sailik Sengupta, Leonard Lausen, Sheng Zha, and Saab Mansour, *Parameter and data efficient continual pre-training for robustness to dialectal variance in arabic*, NeurIPS 2022 Workshop on Efficient Natural Language and Speech Processing (ENLSP), 2022.

- Kihyuk Sohn, *Improved deep metric learning with multi-class n-pair loss objective*, Advances in neural information processing systems **29** (2016).

- Samson Tan, Shafiq Joty, Min-Yen Kan, and Richard Socher, *It's morphin' time! Combating linguistic discrimination with inflectional perturbations*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Online), Association for Computational Linguistics, July 2020, pp. 2920–2935.

- Yu Tanaka, Yugo Murawaki, Daisuke Kawahara, and Sadao Kurohashi, *Building a Japanese typo dataset from Wikipedia's revision history*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (Online), Association for Computational Linguistics, July 2020, pp. 230–236.

- Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler, *Charformer: Fast character transformers via gradient-based subword tokenization*, arXiv preprint arXiv:2106.12672 (2021).

- Emiel van Miltenburg, Chris van der Lee, Thiago Castro-Ferreira, and Emiel Krahmer, *Evaluation rules! on the use of grammars and rule-based systems for NLG evaluation*, Proceedings of the 1st Workshop on Evaluating NLG Evaluation (Online (Dublin, Ireland)), Association for Computational Linguistics, December 2020, pp. 17–27.

- Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel, *ByT5: Towards a token-free future with pre-trained byte-to-byte models*, Transactions of the Association for Computational Linguistics **10** (2022), 291–306.

Weijia Xu, Batool Haider, and Saab Mansour, *End-to-end slot alignment and recognition for cross-lingual NLU*, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (Online), Association for Computational Linguistics, November 2020, pp. 5052–5063.