

Decepticons-IIITH : Morphology-Aware Lexicon Filtering of Back-Translated Data for Low-Resource English-to-Indic Translation

Soujanya Rao*, Sakhil Taneja*, Krishna Chaitanya, Radhika Mamidi

IIIT Hyderabad

(soujanya.rao,sakhil.taneja)@research.iiit.ac.in,
krishnachaitanya.p@students.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract

This study tackles the WMT 2026 Low-Resource Indic Language Translation task, covering English → Assamese, Bodo, Mizo and Nagamese. Low-resource languages lack parallel corpora of sufficient size and quality to train machine translation models. To address this challenge, we propose Morphology-Aware Lexicon Filtering of Back-Translated data to generate good-quality parallel corpora. Initially, monolingual text from publicly available sources is back-translated to English with an NLLB-200-3.3B model fine-tuned on the official training data. To filter this synthetic data, we introduce MALM (Morphology-Aware Lexicon Matching), a filter that checks whether words from a bilingual lexicon are preserved in a translation, using Morfessor stems to handle rich morphology and edit-distance matching to handle borrowed words. For our primary method, training occurs in two stages: fine-tuning on filtered synthetic (back-translated) data followed by fine-tuning on the official parallel data. Whereas for the contrastive method, we are only fine-tuning on the official parallel data. We fine-tune the NLLB-200-3.3B model for Assamese, Mizo, Nagamese and the Indic-Trans2-1B model for Bodo. Our primary systems ranked first among primary submissions on two Category-2 pairs viz. English→Bodo and English→Nagamese. On Nagamese, where only 2,000 official sentence pairs exist, of the two methods we used, the primary system improved by a BLEU score of 8.33.

1 Introduction

Machine translation systems have improved rapidly for languages with huge parallel data, but this progress has not reached the many languages for which such data barely exists; for example, the Indian regional languages illustrate this problem

well. Although they are collectively spoken by hundreds of millions of people, each has only a limited amount of usable parallel corpus. They also differ in script, morphology and the amount of coverage they receive in multilingual pretrained models.

Since parallel data is scarce, we collected monolingual data in the target languages and translated it to the source language to generate synthetic pairs using the back-translation method (Sennrich et al., 2016). However, the effectiveness of this method heavily depends on the quality of the synthetic pairs. A weak back-translation model often becomes the bottleneck due to the low-resource nature of the language, resulting in many translations that drift from the source content. Filtering such data is an open problem. Sentence-embedding models such as LaBSE (Feng et al., 2022) and quality-estimation models such as CometKiw (Rei et al., 2022) are the usual scoring tools, but out of our four languages they only cover Assamese. Simple checks based on language identification and length ratio can detect malformed translations but cannot identify whether the meaning of the translation is preserved.

To address this gap, we propose MALM (Morphology-Aware Lexicon Matching), which scores a synthetic pair by how many words from a bilingual lexicon survive in the translation. The lexicon is induced from the official parallel data with IBM Model 1 (Brown et al., 1993) and then expanded with identity and edit-distance matches, since all four languages borrow English words freely. Matching is done on Morfessor stems (Creutz and Lagus, 2007) rather than surface forms, which keeps the score meaningful for morphologically rich languages, i.e. when a preserved word rarely reappears in its dictionary form. On top of the MALM-filtered synthetic data, our system follows a two-step process: continued fine-tuning of a QLoRA adapter (Hu et al., 2022; Dettmers et al., 2023) on the synthetic pairs, followed by fine-tuning on the official parallel data, with the best check-

*These authors contributed equally and share first authorship.

point selected by chrF++ metric (Popović, 2017). We fine-tune NLLB-200-3.3B (Costa-Jussà et al., 2022) for Assamese, Mizo and Nagamese and IndicTrans2-1B (Gala et al., 2023) for Bodo.

2 Task Setup

The WMT 2026 Shared Task on Low-Resource Indic Language Translation (Pakray et al., 2026) evaluates machine translation systems for English and several low-resource Indic languages. The shared task defines two tracks based on the availability of parallel training data: Category-1 (moderate training data) consists of English $\leftrightarrow$ Assamese, Mizo, Khasi, Manipuri (Bengali) and Manipuri (Meitei); Category-2 (very limited training data) consists of English $\leftrightarrow$ Bodo, Kokborok, Karbi, Nagamese and Tagin.

We participate in four English-to-Indic translation directions, from Category-1 English $\rightarrow$ Assamese (en_to_as) and English $\rightarrow$ Mizo (en_to_lus) and from Category-2 English $\rightarrow$ Bodo (en_to_brx) and English $\rightarrow$ Nagamese (en_to_nag). For each language pair, we submit a primary system trained using MALM-filtered synthetic data followed by fine-tuning on the official parallel corpus and a contrastive system trained only on the official parallel corpus.

3 Dataset

3.1 Official Parallel Data

We train on the official parallel corpora released by the WMT 2026 Low-Resource Indic Machine Translation Shared Task (Pakray et al., 2026). The selected language pairs differ primarily in the amount of available parallel data, with English $\leftrightarrow$ Assamese and English $\leftrightarrow$ Mizo from Category-1, while English $\leftrightarrow$ Bodo and English $\leftrightarrow$ Nagamese from Category-2, as shown in Table 1.

Pair	Pairs	Script
en_to_as	54,000	Bengali-Assamese
en_to_lus	50,000	Latin
en_to_brx	15,214	Devanagari
en_to_nag	2,000	Latin

Table 1: Official parallel training data.

3.2 Monolingual Data

For back-translation, we collect publicly available monolingual data in each target language. For

Assamese, we use the IITB IndicMonoDoc corpus,¹ the indian-nlp Assamese dataset,² and the Assamese portion of the PMIndia monolingual corpus.³ For Bodo, we use the Bodo portions of IITB IndicMonoDoc and AI4Bharat IndicCorpV2.⁴ For Mizo we use the MWire Labs Mizo Language Corpus (4M).⁵ Nagamese has no comparable curated corpus, so we combine the Nagamese sides of publicly available Kaggle and Hugging Face resources: English-Nagamese raw data,⁶ the FNPC-25 parallel corpus,⁷ and the NagaNLP conversational corpus⁸ with news articles we scraped from Nagamese Khobor, a Nagamese-language news site.⁹

3.3 Preprocessing

Monolingual documents are split on line breaks and passed through a uniform cleaning pipeline: zero-width characters are removed, whitespace is collapsed, sentences shorter than 3 or longer than 100 whitespace tokens are dropped and exact duplicates are removed. For Assamese we additionally discard source files containing no Bengali-Assamese script text, which removes a small number of English and Hindi documents from the crawled collections. Where the cleaned pool exceeds our back-

Language	Sentences
Assamese	200,000
Bodo	147,670
Mizo	197,300
Nagamese	88,408

Table 2: Final monolingual data after cleaning.

translation budget of 200,000 sentences, we keep the longest sentences, preferring content-dense text for synthetic-pair generation. The same normalization is applied to the official parallel data before training. Table 2 gives the final monolingual

¹<https://huggingface.co/datasets/cfilt/IITB-IndicMonoDoc>

²<https://github.com/indian-nlp/assamese-dataset>

³<https://data.statmt.org/pmindia/v1/monolingual/>

⁴<https://huggingface.co/datasets/ai4bharat/IndicCorpV2>

⁵<https://huggingface.co/datasets/MWirelabs/mizo-language-corpus-4M>

⁶<https://www.kaggle.com/datasets/belovedorange/english-nagamese-raw-data>

⁷<https://www.kaggle.com/datasets/belovedorange/>

⁸<https://huggingface.co/datasets/agnivamaiti/naganlp-conversational-corpus>

⁹<https://nagamesekhobor.com>

volumes per language; each retained sentence is back-translated in Section 4.1.

4 Methodology

Our pipeline consists of three parts: back-translation of public monolingual text (Section 4.1), filtering of the resulting synthetic pairs with MALM (Section 4.2) and two-stage QLoRA training of the translation models (Section 4.3).

4.1 Back-Translation

For each language pair, we fine-tune the reverse (Indic-to-English) NLLB-200-3.3B model using the official parallel data. NLLB has no language code for Bodo or Nagamese, so their backward models use stand-in source tags: Nagamese reuses the `eng_Latn` token, while the Bodo tag falls outside NLLB’s vocabulary and is treated as an `<unk>` prefix, held fixed across training and inference so that the fine-tuning itself teaches the model its source language. Each fine-tuned model translates the monolingual corpus described in Section 3.2 into English, thereby generating a synthetic parallel sentence pair for each sentence in the monolingual corpus summarized in Table 2. Before any filtering, we validate the raw back-translations with script-independent checks (near-copy rate, English-looking target side, length-ratio distribution) as shown in Table 3. All four languages pass, indicating that the backward models track their sources rather than hallucinate.

Lang	Near copies	LR→En lookalikes	Med. ratio (en/lr)	Trunc (<0.5)
asm	0.03%	0.03%	1.111	7.50%
brx	0.00%	0.00%	1.046	1.97%
lus	1.67%	1.90%	0.808	4.07%
nag	3.63%	0.03%	0.806	5.37%

Table 3: Validation of the raw back-translations: fraction of near-copies, fraction of target sides that look English, median length ratio and fraction of truncated outputs (ratio < 0.5).

4.2 MALM: Morphology-Aware Lexicon Matching

Synthetic data produced by weak backward models is noisy, such noise is known to harm NMT training (Khayrallah and Koehn, 2018). Since embedding and QE (Quality Estimation) based filters cover only Assamese among our languages, we score synthetic

pairs with MALM, which is built from the official parallel data alone in three steps:

1. **Lexicon induction:** We train IBM Model 1 (Brown et al., 1993) for five EM (Expectation–Maximization) iterations in both directions on the official parallel data and keep word pairs scored by the geometric mean of the two directions, restricted to content words. To identify content words, we remove English stopwords and filter frequent target-language tokens appearing in more than 12% of parallel sentences, while excluding tokens shorter than two characters and standalone numerals. We retain pairs occurring in at least three sentence pairs with a score of at least 0.10. This yields seed lexicons of 8,384 (Assamese), 7,189 (Bodo), 5,555 (Mizo) and 750 (Nagamese) entries.
2. **Borrowing-aware expansion:** Low-resource languages borrow English words freely and the seed lexicon misses these because IBM Model 1 can only learn words it sees aligned in the parallel data. We therefore add borrowed words directly, a target-side word that is identical to an English word enters the lexicon with confidence 0.90, and one that differs from an English word by a single character (an adapted spelling) enters with confidence 0.60. This grows the lexicons from 8,384 to 9,277 entries for Assamese, 7,189 to 7,622 for Bodo, 5,555 to 9,149 for Mizo and 750 to 7,907 for Nagamese. The gains are concentrated in the two Latin-script languages, where borrowed English words keep their spelling, it matters most for Nagamese, whose ten-fold expansion makes the filter usable for the pair with only 2,000 parallel sentences.
3. **Morphology-aware scoring:** For each synthetic pair, we look at every English word in the source that appears in our lexicon; each one creates an expectation that its translation shows up in the target sentence. A plain text search would miss most of these, because the target languages attach suffixes to words. We therefore match on stems instead: each word is segmented with unsupervised Morfessor (Creutz and Lagus, 2007), and its longest segment is taken as the stem. An expected translation counts as preserved if its stem matches a stem in the sentence. If stems fail to match,

we fall back to a substring check, i.e. the root only needs to appear inside some word of the sentence, provided it is at least 3 characters long to avoid accidental matches. This fall-back recovers roots fused with case endings in agglutinative Bodo. For example, our Bodo lexicon maps *country* to the inflected form ह्य-दोरनि (“of the country”, with the genitive *-ni*); if the sentence instead contains the locative ह्य-दोराव (“in the country”), no exact match exists, but both forms contain the root ह्यदोर, so the entry still counts as preserved. The MALM score is the fraction of expected words that are found, with each word weighted by its lexicon confidence.

Before the MALM threshold is applied, every pair must pass three basic checks. `fastText` language identification (Joulin et al., 2017) confirms the English side is English, the length ratio between the two sides must fall in $[0.45, 2.2]$ and duplicates are removed. We then keep pairs with a MALM score of at least k , with $k = 0.5$ for Assamese, Mizo and Nagamese. Bodo needs a lower bar, $k = 0.2$, because MALM underestimates it in two ways. Its agglutinative morphology hides roots inside suffixed words, and a lexicon learned from only 15,214 pairs often stores a different synonym than the one the sentence uses. Both push faithful translations to low scores. Two checks confirm the low scores are false negatives of the metric rather than bad translations. First, 98% of Bodo pairs preserve the numbers from their source sentence even within the subset MALM scores as zero. Second, real zero-scored pairs are visibly faithful. The pair “Time8, Kokrajhar, December 7” is translated as “टाइम8, कक्राझार, डिसेम्बर 7”, with both numbers kept and the place name transliterated, yet it scores zero because neither Kokrajhar nor December is a lexicon entry. At $k = 0.5$, only 10,395 Bodo pairs, which forms 7% of the raw synthetic data, would survive. So we relax the threshold rather than discarding good pairs. Table 4 shows the number of sentence pairs retained after filtering.

4.3 Models and Training

We fine-tune NLLB-200-3.3B (Costa-Jussà et al., 2022) for Assamese, Mizo, and Nagamese and IndicTrans2-1B (Gala et al., 2023) for Bodo. During decoding, we use native language tags: `asm_Beng` for Assamese and `lus_Latn` for Mizo under NLLB and `brx_Deva` for Bodo under IndicTrans2-1B.

Lang	k	Before filtering	After filtering	Kept
asm	0.5	200,000	128,939	64.5%
brx	0.2	147,670	47,489	32.2%
lus	0.5	197,300	111,997	56.8%
nag	0.5	88,408	56,343	63.7%
Total	—	633,378	344,768	54.4%

Table 4: Synthetic pairs before and after MALM filtering (robust scorer, expanded lexicon), with the per-language threshold k .

Since Nagamese has no language code in either model, we reuse NLLB’s `eng_Latn` target token as the decoding tag. All training uses 4-bit QLoRA (Dettmers et al., 2023): the base model is quantized to NF4 with double quantization and fp16 compute, and a LoRA adapter (Hu et al., 2022) with rank 32, alpha 64 and dropout 0.1 is placed on the attention projections (q, k, v, out); for IndicTrans2-1B we additionally adapt the feed-forward layers (fc1, fc2).

Training proceeds in two stages, Stage-1 (pre-training) trains the adapter on the MALM-filtered synthetic data with learning rate $3e-4$ for up to 3 epochs for Assamese and Mizo and up to 5 for Bodo and Nagamese, selecting the best epoch by validation loss. Stage-2 (fine-tuning) continues from the Stage-1 adapter on the official parallel data for up to 15 epochs with learning rate $2e-5$ and 100 warmup steps (for Nagamese, up to 25 epochs with learning rate $1e-5$ and warmup ratio 0.1); after each epoch we decode a held-out 5% split and select the best checkpoint by `chrF++` (Popović, 2017), computed with `sacreBLEU` (Post, 2018), with early stopping (patience 5). Both stages use a cosine schedule, weight decay 0.01, paged 8-bit AdamW, gradient checkpointing and an effective batch size of 64 on four RTX 2080 Ti GPUs. The contrastive systems skip Stage-1 and train the adapter on the official parallel data only, under the identical Stage-2 configuration.

5 Results

The organizers evaluated all systems on the blind test sets with BLEU (Papineni et al., 2002; Post, 2018), METEOR (Banerjee and Lavie, 2005), TER (Snover et al., 2006), `chrF++` (Popović, 2017), BERTScore (Zhang et al., 2019) and COMET (Rei et al., 2020). Table 5 reports the official scores of our primary systems (official parallel data plus MALM-filtered back-translation) and our contrastive systems (official parallel data only) on the

Pair	System	BLEU $\uparrow$	METEOR $\uparrow$	TER $\downarrow$	chrF++ $\uparrow$	BERTScore $\uparrow$	COMET $\uparrow$
en_to_as	primary	14.09	35.41	73.78	47.83	85.27	80.59
en_to_as	contrastive	15.00	36.44	72.29	48.98	85.83	81.65
en_to_lus	primary	19.53	46.11	71.37	48.49	81.72	62.45
en_to_lus	contrastive	19.41	46.23	73.81	48.79	81.74	63.45
en_to_brx	primary	16.30	37.56	70.20	50.77	85.25	62.38
en_to_brx	contrastive	15.64	36.18	71.19	49.56	84.78	62.05
en_to_nag	primary	20.34	59.95	76.40	57.66	82.28	69.55
en_to_nag	contrastive	12.01	47.84	105.06	47.78	76.48	62.47

Table 5: Official test-set results for our submissions (Higher is better for all metrics except TER).

four English-to-Indic directions.

Our primary systems rank first among primary submissions in two Category-2 pairs - English-to-Bodo (of 10 primaries), English-to-Nagamese (of 5 primaries). Meanwhile, in Category-1 pairs our primary systems placed 3rd on English-to-Mizo (of 9 primaries) and placed 11th on English-to-Assamese (of 15 primaries).

Since our contrastive system differs from the primary only in the absence of the filtered back-translated data, the primary-contrastive gap in Table 5 directly measures the contribution of MALM-filtered back-translation, and the pattern tracks the size of the official corpus. The largest gains appear on the languages with the least parallel data. On Nagamese, with only 2,000 official pairs, back-translation adds 8.33 BLEU, 9.88 chrF++ and 7.08 COMET, turning the weakest-resourced pair into our strongest submission. Bodo, with 15,214 pairs, improves consistently on all six metrics (+0.66 BLEU, +1.21 chrF++), indicating that the relaxed $k = 0.2$ threshold admitted useful rather than harmful data. On Mizo (50,000 pairs) the effect is mixed, BLEU score improves marginally, TER improves slightly while chrF++ and COMET drop slightly and on Assamese (54,000 pairs) the synthetic data is mildly harmful, costing about one point on most metrics.

The Assamese and Mizo results suggest that once the official corpus is large enough to fine-tune a stronger adapter on its own, synthetic pairs from a weak backward model add more noise rather than improve it. For the Category-2 languages, the Nagamese margin shows that MALM’s borrowing-aware expansion makes the approach viable even when the seed lexicon comes from just 2,000 sentence pairs.

6 Conclusion

We described the Decepticons-IIITH submissions to the WMT 2026 Low-Resource Indic Language Translation task for four English-to-Indic pairs: Assamese, Bodo, Mizo and Nagamese. Our primary system combines back-translation with MALM, a filter that scores synthetic pairs using a bilingual lexicon induced from the official parallel data alone, so it applies even to languages that embedding and quality-estimation models do not cover. The filtered data feeds a two-stage QLoRA process, adapter fine-tuning on synthetic pairs, then fine-tuning on the official corpora.

The main finding from Section 5 is that filtered back-translation helps mainly when training data is scarce. It produced the biggest gains for the two language pairs with the least parallel data, but had little effect once sufficient official data was available. Therefore, Stage-1 synthetic fine-tuning is best reserved for very low-resource languages.

Bodo remains the hardest case for the filter itself. Its agglutinative morphology and small lexicon push many synthetic pairs to lower MALM scores, which we worked around with a relaxed threshold.

Future work includes growing the Bodo lexicon from additional parallel sources, suffix-aware matching in place of unsupervised segmentation alone and a quality-estimation check for Assamese, the one language where such models are usable, a natural further comparison is back-translating Bodo with IndicTrans2, which supports it natively, against our tag-less NLLB backward model. Another comparison can be between models finetuned on unfiltered synthetic data and MALM filtered synthetic data.

References

- Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.
- Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in neural information processing systems*, 36:10088–10115.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 878–891.
- Jay Gala, Pranjal A Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. *arXiv preprint arXiv:2305.16307*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-rank adaptation of large language models**. In *International Conference on Learning Representations*.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. **Bag of tricks for efficient text classification**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Huda Khayrallah and Philipp Koehn. 2018. **On the impact of various types of noise on neural machine translation**. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 74–83, Melbourne, Australia. Association for Computational Linguistics.
- Partha Pakray, Santanu Pal, Advaita Vetagiri, Kshetrimayum Boynao Singh, Sandeep Kumar Dash, Arnab Kumar Maji, Saralin A. Lyngdoh, Lenin Laitonjam, Anupam Jamatia, Ajit Das, Sunita Warjri, Uzzal Sharma, Koj Sambyo, and Riyanka Manna. 2026. Findings of the WMT 2026 shared task on low-resource Indic languages translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Maja Popović. 2017. **chrF++: words helping character n-grams**. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. **A call for clarity in reporting BLEU scores**. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. **COMET: A neural framework for MT evaluation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, et al. 2022. CometKiwi: IST-Unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96.
- Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. **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*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.