No.16 Lost in Translation: Why Zero-Shot Transfer Fails at Detecting Subjectivity in Chinese

李盼嘉乐，计算机系

Background and Aims

Subjective bias (the use of framing, inflammatory adjectives, or opinionated verbs to influence a reader) is an extensive issue in digital media. Neutral Style Transfer has been well studied in English with models such as BERT and GPT, but its effectiveness in debiasing within other languages remains unclear and unexplored.

Current state-of-the-art methods rely on massive, human-annotated parallel corpora (e.g., the Wiki Neutrality Corpus). This creates a bottleneck for non-English languages where such high-quality data is scarce. An open question in Multilingual NLP is whether the semantic concept of "bias" is language-agnostic: **Can a model learn to detect subjectivity in English and apply that knowledge to Chinese without explicit supervision?**

We address this by evaluating the cross-lingual transfer capabilities of mBART-50, a multilingual sequence-to-sequence transformer. We conduct a comparative study of three distinct paradigms to detecting and neutralizing bias in Chinese text:

- **Evaluate Zero-Shot (and Few-Shot) Transfer**: We investigate if mBART's pre-trained multilingual embedding space creates a shared latent representation for "subjectivity," allowing English-trained weights to neutralize Chinese text.
- **Propose Synthetic Data Augmentation**: We introduce a novel training strategy using machine-translated (noisy) parallel data to overcome the lack of native Chinese bias datasets.
- **Analyze the Translation-Preservation Trade-off**: We compare these neural methods against a standard "Pivot Translation" baseline to determine which approach best balances debiasing accuracy with semantic content preservation.

Related works

1. **Neutral Style Transfer & Bias Detection** Research in "Neutral Point of View" (NPOV) editing has focused primarily on English. **Pryzant et al. (2020)** introduced *the Wiki Neutrality Corpus (WNC)* and the first generation of modular editing models, treating bias neutralization as a monolingual rewriting task. Subsequent work by **D'Sa et al. (2021)** applied BERT-based classifiers to detect subjective linguistic markers.
2. **Multilingual Generative Models (mBART) Liu et al. (2020)** proposed **mBART**, a sequence-to-sequence denoising auto-encoder pre-trained on 25 languages. Unlike BERT (which is an encoder only), mBART is designed for generation tasks like translation. While mBART has demonstrated strong **Zero-Shot Transfer** capabilities for objective tasks, its ability to transfer semantics (like debiasing) remains under-explored.
3. **Synthetic Data Augmentation** In low-resource scenarios where parallel training data is unavailable, **Sennrich et al. (2016)** pioneered the use of "Back-Translation" to generate synthetic parallel data. Our work adapts this paradigm to style transfer: we hypothesize that even "noisy" machine-translated data can provide enough semantic signal to train a robust bias neutralizer in Chinese, bypassing the need for expensive human annotation.

Results

Quantitative results of applying fine-tuned models on a testing corpus of 100 manually debiased phrases.

Model	Style Acc (↑)	Content Sim (↑)	Fluency PPL (↓)	Composite Score
Baseline	72%	0.835	36689	0.025
ZeroShot	53%	0.95	23755	0.028
Synthetic	62%	0.964	25694	0.029
FewShot	53%	0.950	23755	0.028

Our baseline model performed the best, while the ZeroShot and FewShot (1% of Chinese data) performed underwhelmingly.

Key References

[1] Pryzant, R., Diehl, R. D., Srivastava, A., & Jurafsky, D. (2020). Automatically neutralizing subjective bias in text. In Proceedings of the AAAI Conference on Artificial Intelligence (pp. 480-489).

[2] Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., ... & Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics, 8, 726-742

Model and Methods

Procuring a testing corpus

Instruction Set 1

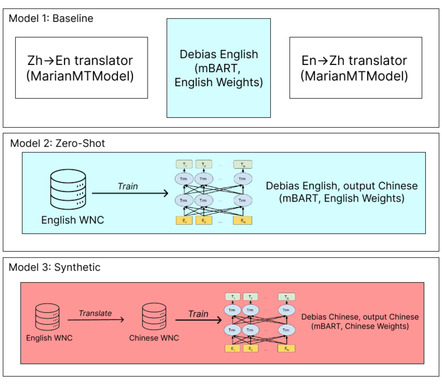
Task 1 (Translate Bias): Look at this Biased English sentence. Translate it into Chinese, but make sure you keep the 'biased' or 'subjective' aspects. If the English calls someone 'radical' or 'infamous,' use a Chinese word that feels equally harsh (e.g., 激进, 臭名昭著).

Instruction Set 2

Task 2 (Human Neutralization): Now, ignore the English. Look at your Chinese translation and rewrite it to be Neutral (like a bored Wikipedia editor). Remove the emotion/judgment words so it states just the facts."

Five participants were provided with 100 randomly selected phrases from the WNC test dataset. Three were provided with Instruction Set 1, and two were provided with Instruction Set 2. This was the final training set used to evaluate each model.

Model Architecture



Sample Model Output

包括瑞安(略微是他们的高年级)的球员在1992年5月帮助俱乐部赢得了青春杯，

Biased Chinese Input

Debias

包括瑞安在内的一批球员于1992年5月帮助俱乐部赢得了青春杯。

Unbiased Chinese Output

Evaluation Criteria

x: The input biased sentence (Source)

$\hat{y}$: The generated neutral sentence (Target)

$f_{\theta_{bert}}$: A fine-tuned BERT binary classifier for neutrality detection.

$E(\cdot)$: The sentence embedding function (LaBSE).

P_{LM}: A pre-trained Language Model (GPT-2 Chinese) for probability estimation.

1. Style Transfer Accuracy (ACC)

$$ACC = \frac{1}{N}\sum_{i=1}^{N} I\left[f_{\theta_{bert}}(\hat{y}_i) = \text{Neutral}\right]$$

2. Semantic Content Preservation (SIM)

$$SIM(x, \hat{y}) = \cos\left(v_x, v_{\hat{y}}\right) = \frac{v_x \cdot v_{\hat{y}}}{|v_x||v_{\hat{y}}|}$$

3. Fluency / Perplexity (PPL)

$$PPL(\hat{y}) = \exp\left(-\frac{1}{L}\sum_{t=1}^{L} \ln P_{LM}\left(w_t | w_{<t}\right)\right)$$

4. Composite Score (Joint Metric)

$$S_{comp} = \sqrt[3]{ACC \times SIM \times \frac{1}{\log(PPL)}}$$