Improving Chinese-English Translation of Conversational Text via a Two-Stage Pipeline

UC Berkeley W266 Final Report – Summer 2025

Huibin Chang

Abstract

This project investigates the challenge of translating informal, pronoun-dropping Chinese messages into English using large neural machine translation models. I construct and evaluate a multi-step pipeline that enhances translation by incorporating speaker-aware pronoun recovery and fine-tuning. Using the BOLT Chinese-English SMS/Chat corpus, I benchmark baseline and fine-tuned mBART models, implement an enhanced rule-based pronoun recovery module, and explore back-translation as an additional source of signal. I evaluate translation quality using BLEU, BLEURT, and COMET, and perform qualitative error analysis. The results show that even simple recovery methods can improve downstream translation when combined with targeted fine-tuning.

1. Introduction

Low referential density languages such as Chinese, Japanese, and Korean present a challenge to machine translation (MT) systems due to their frequent omission of pronouns in informal contexts, a phenomenon known as pro-drop. When translating to English, these omitted subjects or objects often lead to ambiguity or error. While previous research has tackled pronoun recovery as a separate NLP task, integration into full translation pipelines remains underdeveloped. My original proposal centered on this issue, but through the course of experimentation, the focus expanded to the broader challenge of translating low-resource, noisy, conversational Chinese (e.g., SMS, chat messages), which shares many of the same ambiguities.

I pose the question: *Can we improve translation of colloquial Chinese to English through an explicit pronoun recovery step, and can this improvement be captured via fine-tuning large pre-trained models?*

2. Data and Related Work

I use the BOLT Chinese-English SMS/Chat Parallel Corpus[1] (Tracey et al., 2021) from the Linguistic Data Consortium (LDC). This dataset contains turn-by-turn annotated SMS conversations, with speaker tags and human English translations. It is a rich but noisy resource: source sentences (Chinese) are short, informal, ungrammatical, and context-dependent, much like a distinct language register. On the other

[1] Descriptions of the data are provided here: https://catalog.ldc.upenn.edu/LDC2021T11. The data is downloaded from the UCB library: https://digitalassets-lib-berkeley-edu.libproxy.berkeley.edu/UCBonly/ldc/2021T11/.

hand, the data contains human-translated target sentences[2], speaker ID, which can be used to augment the quality of source sentences.

Prior work falls into two categories: 1) Pronoun recovery as a preprocessing task (Wang et al., 2023): aligning English and Chinese to heuristically recover zero pronouns. 2) Architectural innovations (Ri et al., 2021; Yang et al., 2019): integrating discourse or document-level context into translation models. Wang et. al. (2023) provide a helpful survey. This work contributes to the first category by introducing a simple, robust, and speaker-aware recovery module, and partially to the second by exploring multi-stage fine-tuning[3].

3. Methodology

3.1 Overview

The central hypothesis is that restoring dropped pronouns in input (e.g., Chinese) improves translation accuracy, especially when models are fine-tuned on such restored input. The pipeline consists of three stages: 1) pronoun recovery (data augmentation), 2) translation modeling (fine-tuning mBART), and 3) evaluation across multiple test sets using BLEU, BLEURT, COMET..

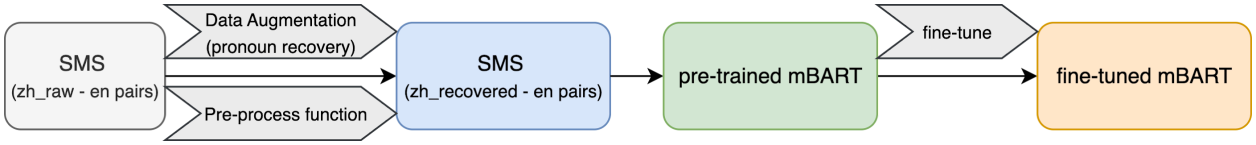
3.2 Pronoun Recovery

To address the pro-drop phenomenon in colloquial Chinese, I implemented a rule-based function that enriches the raw Chinese text with likely omitted subject pronouns, using target English references as a guide. Multiple sets of rules with varying degrees of complexity were experimented and then manually checked for quality, and for this project I settled on a set of relatively simple rules: If the target sentence starts with "I", "I'll", "I'm", etc., *and* the Chinese sentence lacks a pronoun, I prepend “我” (I). Similar but more complex rules are applied to missing second-person pronouns. For example, 6.7% of 16,000 training examples have “我” (I) added to the raw Chinese sentence, and 2.8% have “你” (you) added. Among 500 holdout examples, 8% have 我” added and 3.2% have “你” added.

3.3 Model Architecture

The base model is facebook/mbart-large-50-many-to-many-mmt, an encoder-decoder transformer pre-trained on multiple languages, and the base model is used in two ways: 1) Base inference (zero-shot): tested directly on raw source sentences and refined source sentences; 2) Fine-tuned: trained on refined source → target pairs using a custom training loop. This pipeline is shown in Figure 1.

Figure 1: Two-stage Fine-tuning pipeline using BOLT SMS data



[2] Visual inspection reveals that some of the human-translated target sentences are erroneous.
[3] The backbone model is `facebook/mbart-large-50-many-to-many-mmt`, a multilingual sequence-to-sequence model.

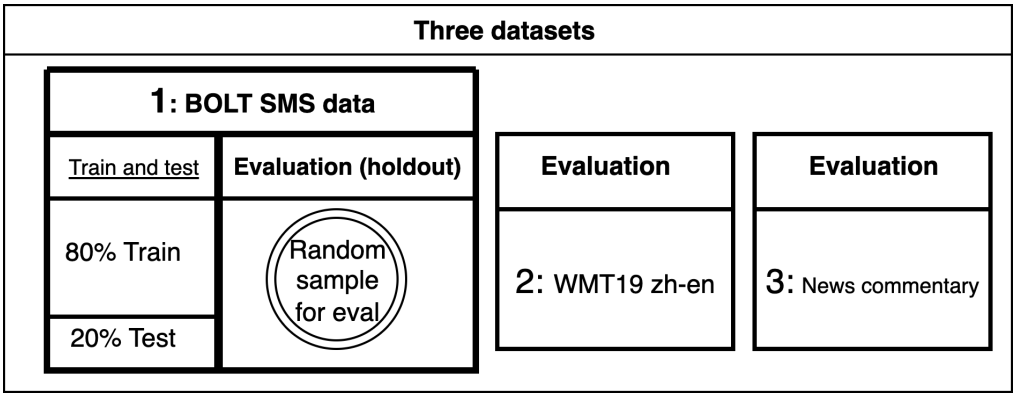
3.6 Fine-tuning considerations and datasets

Several key considerations for fine-tuning and experiments should be noted. While the original data contains more than 5500 pairs of source and target files, and each pair of files contains dozens of sentences (different turns in a conversation), totalling in almost 200,000 sentence pairs, experimenting with the hardware constraint (Colab Pro+) resulted in a pipeline using only 20,000 sentence pairs (80% training, 20% validation), short input and output generation lengths (16 tokens), and limited number of experiments conducted prior to this report.

I use three markedly different datasets for evaluation (Figure 2). The main dataset is the SMS dataset, which, as mentioned above, has more than can be processed for training. I therefore use approximately one-tenth of it (20,000 sentence pairs) for training and testing (an 80-20 split), and randomly select 500 sentence pairs from the remaining SMS data. A benefit of this approach of randomly selecting evaluation examples from a large pool is the ability to conduct repeated evaluations on different random samples, thereby mitigating the likelihood of obtaining extreme evaluation scores by chance.

The second dataset is WMT19, and the third is the HuggingFace News Commentary data. Both are used only for evaluation.

Figure 2: Three datasets for training, testing, and experiments (evaluation)



3.7 Experiments and evaluation metrics

After the mBART model is fine-tuned on improved Chinese and English sentence pairs, I compare the performance of the two models (baseline and fine-tuned) on three metrics (BLEU, BLEURT, and COMET) across three tasks/datasets, as shown in Table 1.

1. The first evaluation task is to translate the BOLT SMS Chinese to English and compare the translated English with the English references. For this task, the baseline model translated both raw Chinese and improved Chinese separately, whereas the fine-tuned model only translated from improved Chinese to English.
2. The second evaluation task is to translate the WMT19 English-Chinese data.
3. The final evaluation task is to translate the Hugging Face News Commentary data.

Table 1: Experiment datasets, tasks, and metrics

Experiment settings			Model	
			Baseline	Fine-tuned
Evaluation: data and task	SMS	Raw Chinese-English	BLEU, BLUERT, COMET	---
	SMS	Improved Chinese-English	BLEU, BLUERT, COMET	BLEU, BLUERT, COMET
	WMT19		BLEU, BLUERT, COMET	BLEU, BLUERT, COMET
	News Commentary		BLEU, BLUERT, COMET	BLEU, BLUERT, COMET

The purpose of using multiple characteristically different datasets to evaluate is to determine whether fine-tuning causes catastrophic forgetting. As mentioned above, because the SMS data contains short, ungrammatical, and context-dependent casual conversations, one must guard against the scenario where an idiosyncratic dataset leads to a model that is ungeneralizable.

4. Results discussion and error analysis

4.1 Results discussion

The results in Table 2 show a consistent and meaningful improvement from the fine-tuned model across all metrics when evaluated on the holdout set of unseen SMS data. Specifically, the fine-tuned model outperforms the baseline model on BLEU, BLEURT, and COMET, confirming that augmenting the training data with a simple speaker-aware pronoun recovery procedure has downstream benefits. The COMET score improves by more than 0.28 (from 0.704 to 0.732), which is especially notable since COMET is sensitive to semantic adequacy and contextual fluency.

Table 2: Evaluation

Dataset	Size (pairs)	Model - Task	BLEU (Avg)	BLEURT	COMET
BOLT SMS Holdout	500	Base (raw_zh - en)	0.103	-0.073	0.702
		Base (improved_zh - en)	0.107	-0.058	0.704
		Fine-tuned (improved_zh - en)	**0.132**	**0.055**	**0.732**
WMT19	3,981	Base Model WMT19	0.164	0.015	**0.776**
		Fine-tuned Model WMT19	**0.166**	**0.044**	0.773
News Commentary	69,206	Base Model	0.056	**-0.220**	**0.641**
		Fine-tuned Model	**0.056**	-0.223	0.638

Importantly, this gain does not come at the cost of generalization. On out-of-domain evaluation sets, i.e., WMT19 and News Commentary, the fine-tuned model performs comparably to the baseline model. On WMT19, the fine-tuned model only slightly underperforms the base model on COMET, but improves marginally on BLEU and BLEURT. On the News Commentary set, the differences between the base and fine-tuned models are small across all three metrics. Together, these results suggest a mild but welcome

Pareto improvement: fine-tuning on pronoun-recovered conversational Chinese improves performance on the target domain (SMS), while largely preserving or even improving generalization to more formal domains.

4.2 When and why does the model's performance degrade?

Further experiments revealed a training dynamic that aligns with intuition: when the model is trained for more epochs, or when the learning rate is greater, it continues to improve on SMS translations (higher BLEU, BLEURT, and COMET scores on the SMS holdout), but its performance on WMT19 and News Commentary begins to degrade. This points to a tradeoff between domain adaptation and generalization. Given that the fine-tuning corpus is stylistically and structurally different from WMT-style data, overfitting to short, informal utterances may hurt generalization to longer, more grammatical input.

For example, many sentences in the training data are two-character utterances whose meaning deviates from their usual ones and may only be vaguely determined by the conversational context or ephemeral cultural references. While memorization of such examples can help the model's performance on similar datasets from the same time period, such learning is likely to hinder the model’s general ability to perform MT. Some example sentence pairs from the training and evaluation datasets, as well as model translations, are shown in the Appendix (section A.3 and Table A2, where concrete examples are analyzed).

In fact, another likely explanation for the model’s relatively weaker performance on WMT19 and News Commentary is that the training pipeline was constrained to short input lengths (translation maximum 32 tokens). This decision, driven by hardware limitations, reduces the model’s ability to capture long-distance dependencies — a key feature of formal text. Thus, the results reflect a structural bottleneck rather than a modeling limitation.

Overall, these findings demonstrate the benefit of pronoun recovery and targeted fine-tuning in noisy, low-resource domains without significantly sacrificing broader translation quality.

5. Conclusion and Future Work

When we use short, context-dependent, and ungrammatical conversation texts to train models, it might lead to garbage-in-garbage-out results. But this project demonstrates that even lightweight, rule-based methods for recovering raw text guided by reference translations and speaker roles can improve the Chinese-to-English translation of conversational text. Fine-tuning on such augmented data significantly improves translation performance on unseen SMS content, while generalizing well to out-of-domain corpora. This represents a modest but meaningful Pareto improvement: gains in the target domain without sacrificing broader translation robustness.

The results validate the hypothesis that pre-processing techniques tailored to the linguistic and pragmatic characteristics of the source domain can enhance downstream translation. More broadly, this work offers a replicable framework for adapting general-purpose translation models to low-resource, informal, or context-dependent registers.

Several extensions could build on this foundation. The first is to develop a more advanced general raw text improvement function. The current rule-based recovery module could be enhanced using named entity recognition (NER) and syntactic parsing to detect dropped entities with higher precision. Such techniques could be recruited as a standalone tool or a helpful preprocessing stage in a larger NLP pipeline. Another promising direction is to develop learned recovery models. For example, a natural next step for this project is to train a dedicated model to perform source language sentence recovery using mT5 or BART. This learned encoder-decoder stage would generalize the source text recovery pipeline beyond rigid heuristics.

References

Tracey, J., Delgado, D., Chen, S., & Strassel, S. (2021). *BOLT Chinese SMS/Chat Parallel Training Data* (p. 116628 KB) [Dataset]. Linguistic Data Consortium. https://doi.org/10.35111/CV5N-E349

Wang, K., Zhao, X., Li, Y., & Peng, W. (2023). PROSE: A Pronoun Omission Solution for Chinese-English Spoken Language Translation. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2297–2311). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.141

Wang, L., Liu, S., Xu, M., Song, L., Shi, S., & Tu, Z. (2023). *A Survey on Zero Pronoun Translation* (arXiv:2305.10196). arXiv. https://doi.org/10.48550/arXiv.2305.10196

Yang, J., Tong, J., Li, S., Gao, S., Guo, J., & Xue, N. (2019). Recovering dropped pronouns in Chinese conversations via modeling their referents. In J. Burstein, C. Doran, & T. Solorio (Eds.), *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. 892–901). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1095

Appendix

A.1 Operational details

A.1.1 Raw data processing

The raw data is stored as .xml files in two folders: /source/ and /target/. Each file contains one conversation with multiple sentence pairs. The files in /source/ and /target/ can be matched using the

prefix portion of the file names (and a manual examination of a large number of processed data is performed to ensure that the source and target sentences are matched). The conversation-level matched pairs are then broken down, using speaker and turn IDs, to create sentence-level matched pairs. Since parsing the .xml files is time-consuming, I save the parsed data locally in a .csv file for repeated use.

A.1.2 Tokenizer and model selection

With experimentation, the "facebook/mbart-large-50-many-to-many-mmt" system is chosen because the pre-trained version performs better on the SMS dataset than some experimented T4 and Helsinki-NLP/OPUS-MT models. "Max_length" is set to 16 in the tokenizer due to compute constraints.

A.1.3 Hardware and training

The fine-tuned model is trained on Colab's T4 High-RAM runtime. Several experiments were conducted to determine training parameters that would not cause the system to crash. Increasing the number of epochs and/or the learning rate would lead to better training and evaluation outcomes on the SMS data, albeit at the cost of degraded performance on out-of-domain data (WMT19 and News Commentary).

A.2 Sample data and augmented input

Table A1: Sample data and improved Chinese input

conversation_id	turn_id	participant	speaker	zh_raw	en	zh_recovered
Examples that the pre-processing stage correctly improved the raw Chinese sentence						
CHT_CMN_20130121.0014	s44	133655	A	饿死了	I'm starving	我饿死了
CHT_CMN_20120704.0000	s23	133489	A	要买500块以上的电饭煲才行吗	I should only buy an electric rice cooker for more than 500 yuan, right?	我要买500块以上的电饭煲才行吗
CHT_CMN_20130519.0003	s76	134219	A	暑假要继续教育[表情]	I ought to continue studying in the summer holidays [emoticon]	我暑假要继续教育[表情]
CHT_CMN_20120811.0010	s40	135882	A	不是说还有个卖家要买电视吗	Didn't you say there is another seller wanting to buy a television?	你不是说还有个卖家要买电视吗
Examples that the pre-processing stage incorrectly improved the raw Chinese sentence						
CHT_CMN_20120905.0005	s49	131944	B	是你帮查的？	Did you help look it up?	你是你帮查的？
CHT_CMN_20130701.0000	s30	135610	B	以后就是你甩人了	You'll be the one dumping others in future	你以后就是你甩人了
Examples that the pre-processing does nothing to an example pair (the majority of the data)						
CHT_CMN_20110513.0002	s141	131525	A	我把琴带来	I'll bring over my instrument	我把琴带来

In Table A1, Column "zh_recovered" is the output of the data augmentation. Most examples are unchanged (88.8% of the holdout evaluation dataset). Among augmented examples, most are correct, and some are incorrect. It can be seen in Table A1, English references (human translation) are sometimes grammatically or semantically flawed.

A.3 Sample model translation

Table A2 shows 3 examples where the model translates better than the reference and 3 examples where the model did not correctly translate. In the incorrect examples, the first 2 source sentences are context-dependent, and the colloquial use case conveys a very different meaning from translating each individual token. In the third example, where the model performs poorly, the adjective's meaning (literally meaning mind not being present) is altered somewhat by the objective (killers). The model translated the sentence somewhat literally and missed the subtle change in the meaning of the subjective.

Table A2: Translations by fine-tuned mBART model

Dataset	Chinese	English reference	Model translation
Model translates better than reference (author's judgement)			
Holdout	张帆什么时候去巴西的？	When is Zhang Fan going to Brazil?	When did Zhang Han go to Brazil?
WMT19 Validation	就连街头人士也这么认为	And the guy on the corner street, too	Even people on the street think so.
News Commentary	一位和平自由的中国先锋	A Chinese Champion of Peace and Freedom	A pioneer of peace and freedom in China
Model mistranslates			
Holdout	局里你走走关系吧	Find some relations in the bureau	In the office, you can go
WMT19 Validation	这是展示这个球场的最佳舞台	This is a great showcase for this course	This is the best stage to showcase the stadium.
News Commentary	心不在焉的杀戮者	Absent-Minded Killers	Killers without mind