Cultural Adaptation of Recipes

Cao, Yong; Kementchedjhieva, Yova; Cui, Ruixiang; Karamolegkou, Antonia; Zhou, Li; Dare, Megan; Donatelli, Lucia; Hershcovich, Daniel

Published in:
Transactions of the Association for Computational Linguistics

DOI:
10.1162/tacl_a_00634

Publication date:
2024

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Cultural Adaptation of Recipes

Yong Cao\textsuperscript{1,2,4}, Yova Kementchedjhieva\textsuperscript{2}, Ruixiang Cui\textsuperscript{2}, Antonia Karamolegkou\textsuperscript{2}, Li Zhou\textsuperscript{2,3}, Megan Dare\textsuperscript{4}, Lucia Donatelli\textsuperscript{4} and Daniel Hershcovich\textsuperscript{2}

\textsuperscript{1}Huazhong University of Science and Technology, China
\textsuperscript{2}Department of Computer Science, University of Copenhagen, Denmark
\textsuperscript{3}University of Electronic Science and Technology of China, China
\textsuperscript{4}Department of Language Science and Technology, Saarland University, Denmark

\{yongcao, yova, rc, antka, dh\}@di.ku.dk, li.zhou@uestc.edu.cn,
\{mdare, donatelli\}@coli.uni-saarland.de

Abstract

Building upon the considerable advances in Large Language Models (LLMs), we are now equipped to address more sophisticated tasks demanding a nuanced understanding of cross-cultural contexts. A key example is recipe adaptation, which goes beyond simple translation to include a grasp of ingredients, culinary techniques, and dietary preferences specific to a given culture. We introduce a new task involving the translation and cultural adaptation of recipes between Chinese- and English-speaking cuisines. To support this investigation, we present CulturalRecipes, a unique dataset composed of automatically paired recipes written in Mandarin Chinese and English. This dataset is further enriched with a human-written and curated test set. In this intricate task of cross-cultural recipe adaptation, we evaluate the performance of various methods, including GPT-4 and other LLMs, traditional machine translation, and information retrieval techniques. Our comprehensive analysis includes both automatic and human evaluation metrics. While GPT-4 exhibits impressive abilities in adapting Chinese recipes into English, it still lags behind human expertise when translating English recipes into Chinese. This underscores the multifaceted nature of cultural adaptations. We anticipate that these insights will significantly contribute to future research on culturally aware language models and their practical application in culturally diverse contexts.

1 Introduction

Cooking recipes are a distinct form of procedural text whose accurate interpretation depends on several factors. Familiarity with ingredients and measurement units, common sense about the cooking environment, and reasoning about how tools and actions affect intermediate products in the cooking process are necessary to successfully craft a recipe. Such knowledge varies by culture and language, as a result of geography, history, climate, and economy (Albala, 2012). These factors impact the frequency of ingredient usage, the available forms and cost of heat for cooking, common taste profiles, written recipe style, etc. \textsuperscript{(2)}.

Identifying and adapting to cultural differences in language use is important and challenging (Hershcovich et al., 2022). Recipe translations with current machine translation technology may gloss over culture-specific phraseology or yield mistranslations due to a lack of grounding in the physical and cultural space. Literal translations are often opaque or odd: a Chinese dish, 夫妻肺片 (literally, ‘husband and wife lung slices’), can be adapted in translation to ‘Sliced Beef in Chili Sauce’ for English-speaking cooks. Structural patterns in recipes in different cultures (e.g., \textit{mise en place}\textsuperscript{1}) additionally make straightforward recipe translation difficult: cuisines differ in dish preparation methods, and temporal dependencies between actions complicate the disentanglement of recipe actions (Kiddon et al., 2015; Yamakata et al., 2017).

In this work, we introduce the task of adapting cooking recipes across languages and cultures. Beyond direct translation, this requires adaptation with respect to style, ingredients, measurement units, tools, techniques, and action order preferences. Focusing on recipes in Chinese and English, we automatically match pairs of recipes for the same dish drawn from two monolingual corpora,

\begin{itemize}
\item \textsuperscript{1}In French cooking, \textit{mise en place} is the practice of measuring out and cutting all ingredients in advance.
\end{itemize}

Action Editor: Taro Watanabe. Submission batch: 3/2023; Revision batch: 10/2023; Published 1/2024.
© 2024 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.
and train text generation models on these pairs. We evaluate our methodology with human judgments and a suite of automatic evaluations on a gold standard test set that we construct. We provide ample evidence that recipe adaptation amounts to more than mere translation and find that models finetuned on our dataset can generate grammatical, correct, and faithful recipes, appropriately adapted across cultures. Intriguingly, Large Language Models (LLMs) outperform our finetuned models in both automatic and human evaluations, even without training on our paired dataset. This unexpected result opens multiple avenues for future research, including how large-scale pre-training could complement our dataset and nuanced evaluation metrics that could better capture the complexities of recipe adaptation. Our contributions are as follows:

(a) We introduce the task of cross-cultural recipe adaptation and build a bidirectional Chinese-English dataset for it, CulturalRecipes (§3).

(b) We experiment with various sequence-to-sequence approaches to adapt the recipes, including machine translation models and multilingual language models (§6).

(c) We evaluate and analyze the differences between Chinese- and English-speaking cultures as reflected in the subcorpora (§4) and to the translation and adaptation of recipes (§6).

Our dataset, code, and trained models are available at https://github.com/coastalcph/cultural-recipes.

2 Cultural Differences in Recipes

Extensive cross-cultural culinary research reveals compelling differences in ingredients, measurement units, tools, and actions, each reflecting historical, geographical, and economic influences unique to each culture (Albala, 2012). For example, the historical reliance on open flame cooking in China has cultivated an array of oil-based cooking techniques exclusive to Chinese cuisine. Further complexities arise from culture-specific terminologies for cooking methods and dish names, which pose formidable challenges to translation and adaptation (Rebechi and da Silva, 2017). Additionally, the visual presentation of online recipes exhibits striking contrasts across different cultural contexts (Zhang et al., 2019a). Delving deeper, culinary preferences also demonstrate regional patterns in flavor profiles; Western cuisines tend to combine ingredients that share numerous flavor compounds, while East Asian cuisines often intentionally avoid such shared compounds (Ahn et al., 2011). These intricate cultural nuances underscore the complexity and diversity inherent in global culinary practices, thereby emphasizing the intricacy involved in adapting recipes across different cultures.

Examples. Figure 1 presents a Mandarin Chinese recipe and its human-authored adaptation to American English, highlighting key differences:

1. Ingredients. Distinct ingredients feature prominently in each recipe; the Chinese version highlights 米酒 ‘rice wine’, 红豆 ‘red beans’, and 姜皮老姜 ‘ginger with skin’. Interestingly, while ‘red bean’ is referenced in Chinese recipes, the equivalent ingredient is typically recognized as ‘adzuki beans’ in Western countries.

2. Measurement units. Chinese recipes often rely on imprecise measurements, guided by the cook’s experience, while American English recipes use precise U.S. customary or Imperial units like ‘cups’, ‘inches’, ‘pints’, and ‘quarts’. Occasion-
(3) **Tools.** Specificity varies between recipes, with English recipes typically specifying pot sizes while Chinese recipes provide more general descriptions. Chinese recipes also favor stovetop cooking over ovens, contrasting with their English counterparts.

(4) **Actions by cook.** Preparation methods often vary between Chinese and English recipes. For instance, Chinese recipes usually involve shredding ginger, while English recipes recommend peeling and juliennining. Additionally, unique processes like 水焯 'blanching', common in Chinese cooking to remove unwanted flavors, are rarely found in English recipes. These differences highlight the subtle cultural nuances in similar recipes.

**Over-generalization and Bias.** In a study of cultural adaptation, it is important to recognize that the concept of ‘‘culture’’ is multifaceted and complex. When we refer to Chinese- and English-speaking cultures throughout this work, we make the simplifying assumption that there are general features that characterize the cooking of these cultures and make them distinct in certain systematic ways. We recognize that there is enormous diversity within these simplistic categories,2 but as a first step towards the adaptation of recipes across cultures, we restrict ourselves to the coarse-grained level only.

To enable the development and benchmarking of recipe adaptation, we build a dataset for the task.

3 The CulturalRecipes Dataset

Our dataset, CulturalRecipes, builds on two existing large-scale recipe corpora in English and Chinese, respectively. We create two collections of automatically paired recipes, one for each direction of adaptation (English→Chinese and Chinese→English), which we use for training and validation in our recipe adaptation experiments (§6). Additionally, CulturalRecipes incorporates a small test set of human adaptations expressly crafted for the task in each direction, serving as references in our experimental evaluation.

3.1 Recipe Corpora

We source recipes from two monolingual corpora: RecipeNLG (Bień et al., 2020) and XiaChuFang (Liu et al., 2022).3 RecipeNLG consists of over 2M English cooking recipes. It is an extension of RECIPE1M (Salvador et al., 2017) and RECIPE1M+ (Marin et al., 2019), with improvements in data quality. XiaChuFang consists of 1.5M recipes from the Chinese recipe website xiachufang.com, split into a training and evaluation set.

We use the training set and clean it by removing emojis,4 special symbols, and empty fields. We use the title, ingredients, and cooking steps fields of the recipes from both corpora. The recipes in RecipeNLG consist of nine ingredients and seven steps on average, and in XiaChuFang, of seven ingredients and seven steps. As these two corpora are independent and monolingual, discovering recipe equivalents between them is not trivial.

3.2 Recipe Matching Rationale

Our recipe matching procedure relies on the following assumption: If two recipes have the same title, they describe the same dish. This assumption can be applied even in a monolingual context: if two recipes are both titled ‘Veggie Lasagna’, we can assume that they describe the same dish (Lin et al., 2020; Donatelli et al., 2021). It is permissible that there is some mismatch in the set of ingredients, in the number and sequence of steps, in the measurement units and exact amounts, etc. The same assumption can be said to hold for a recipe with a slightly different, but semantically equivalent title, e.g., ‘Vegetable Lasagna’. Similarly, if we take the Chinese recipe title 卷心菜番茄牛肉汤, we translate it to ‘Cabbage tomato beef soup’ and we find a recipe with a very similar title in English, e.g., ‘Cabbage beef soup’, we can assume that these two recipes describe the same dish. The degree to which this assumption holds depends on the quality of translation of recipe titles from one language into the other, on the measure of similarity, and on how much distance we allow for between two recipe titles before they are no longer considered semantically equivalent. These factors guide our approach to building a silver-standard dataset for the task, further described below, with the procedure also

---

2For example, southern and northern Chinese cuisines are vastly different, with rice and wheat as staples, respectively.


4Despite their potential significance, we remove emojis since they occur only in a few XiaChuFang recipes.
Figure 2: Training and validation (left) and test (right) silver-standard data compilation in the direction Chinese→English. The process is analogous for the opposite direction.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Mean # Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>zh→en</td>
<td>44,5k</td>
<td>159.1</td>
</tr>
<tr>
<td>en→zh</td>
<td>43,8k</td>
<td>117.1</td>
</tr>
</tbody>
</table>

Table 1: Statistics of (many-to-many) training, (one-to-one) silver-standard and gold-standard (human-written) evaluation sets for both directions. zh: Chinese. en: English. We count tokens with whitespace tokenization for English and jieba text segmentation for Chinese.

3.3 Silver-standard Data

Training and Validation Sets. We obtain training recipe pairs by (1) automatically translating all recipe titles in the Chinese corpus to English using a pre-trained machine translation model (Tiedemann and Thottingal, 2020); (2) encoding all English and translated Chinese titles with the MPNet sentence encoder (Song et al., 2020) to obtain two embedding spaces; and (3) in each direction (English→Chinese and Chinese→English), retrieving up to $k = 10$ nearest neighbors per source title from the target space, and filtering out any neighbors that have a cosine similarity against the source title lower than 0.85. The resulting sets, one in each direction, contain multiple reference targets for each source recipe. We further split the matches into training and validation sets.

We recognize that the aforementioned procedure can be susceptible to various sources of noise due to the translation of titles, the encoder representations, and the fixed similarity threshold. We trust that the signal-to-noise ratio should still be sufficient to enable model learning, but for evaluation we need cleaner, more representative data.

Test Set. We are able to eliminate one of the aforementioned sources of noise by collecting manual translations of Chinese recipe titles into English and vice versa from websites that explicitly mention the original dish name when presenting an adapted version. This should resolve issues like being translated literally by an automatic MT system (see §1). To supplement these titles with a corresponding list of ingredients and steps, we look up each title in the recipe corpus of the corresponding language and find the most similar title within, allowing for different capitalization, punctuation and slight differences in word choice and order, e.g., ‘Rice with caramelized leeks’ and ‘Caramelized Leek Rice’ (we manually inspect candidate matches to ensure semantic equivalence).

The resulting test set closely resembles the training data, thus allowing us to determine how well the models we train do in the setting they were trained for (mapping between automatically matched recipes). In order to evaluate the models’ ability to perform the true task we want to solve, i.e. adapting specific recipes from one culture to another, we also construct a gold-standard test set.

---

5Prior to the procedure described below, we filter out recipes longer than 512 subword tokens (arbitrarily using the mT5 tokenizer; Xue et al., 2021) to facilitate using the neural approaches described in §6.

6Helsinki-NLP/opus-mt-zh-en.

7sentence-transformers/all-mpnet-base-v2.
Figure 3: Screenshot from our human recipe adaptation platform, demonstrating the English→Chinese direction, with the source recipe on the left. On the right, participants should adapt the title, ingredients, and steps based on their culinary knowledge and cultural habits.

3.4 Gold-standard Test Data

We include human-written adaptations in our dataset as the ground truth for reference-based evaluations (§5.1, §5.2) and as a point of comparison in human evaluations (§5.3). We select 41 English recipes and 25 Chinese recipes manually from the silver test sets to adapt each to the other culture.

We develop an in-house web application as our recipe writing platform, illustrated in Figure 3. Our guidelines encourage participants to adapt recipes based on their culinary knowledge and cultural customs. We give participants the option to skip a recipe if they are not able to confidently adapt it. Six native Chinese speakers proficient in English with experience in both Chinese and Western cooking volunteered for the task, spending 6.4 minutes on average to adapt a recipe. Subsequently, three of the authors, fluent in both English and Chinese, who have substantial cooking experience, hand-corrected and improved all adapted recipes, including filtering incomplete source recipes, and correcting grammatical errors, spelling mistakes, and non-executable recipe expressions.

4 Corpus Analysis

Here, we perform a data-driven analysis to investigate how the cultural differences discussed in §2 are realized in English and Chinese recipe corpora through the lens of distributional semantics.

4.1 Embedding Alignment

In this analysis, we train static monolingual word embeddings on English and Chinese recipe data, respectively, as a means of capturing their distributional properties. While the global geometry of English and Chinese distributional spaces is similar (Lample et al., 2018), we hypothesize that cultural differences would lead to mismatches in the local geometry of the two spaces (Søgaard et al., 2018). We test this hypothesis through cross-lingual embedding alignment, wherein the English and Chinese embeddings are aligned through a linear mapping to obtain a cross-lingual embedding space, in which semantic equivalents between the two languages should occupy a similar position.

We train monolingual word embeddings using Word2Vec based on a skipgram model by Mikolov et al. (2013b) on the entire English and Chinese corpora (§3.3),10 and align them using VecMap (Artetxe et al., 2017) with weak supervision from a seed dictionary of 15 culturally neutral word pairs we manually curate.11

4.2 Analysis

We use the top 100 most common Chinese content words in the XiaChuFang dataset (not included in our seed dictionary) as query terms and retrieve their five nearest neighbors in the English embedding space, thus inducing a bilingual lexicon from the cross-lingual embedding space (Mikolov et al., 2013a). We manually evaluate this dictionary for correct literal translations and report performance in terms of Precision@5: The ratio of query words for which the correct translation is among the word’s five nearest neighbors in the target space (Lample et al., 2018). The equation is defined as:

\[
\text{Precision}@k = \frac{N@k}{N}
\]

where \(N@k\) is the number of pairs with the correct literal translation in top \(k\) nearest neighbors and \(N\) is the total number of pairs.

The result is 68% (i.e., 68 of 100 query words were correctly mapped), which indicates that (a)

10We train 300-dimensional embeddings for 5 epochs using a minimum frequency count of 10, window size of 5, and 10 negative samples. Chinese text is tokenized with jieba.
Table 2: Top-5 examples from bilingual lexicon induction with underlined literal matches, mismatches, and matches that can be attributed to cultural differences.

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>Nearest Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>水果</td>
<td>fruit</td>
<td>fruit, fruits, kiwi, strawberry, seasonal</td>
</tr>
<tr>
<td>沙拉</td>
<td>salad</td>
<td>feta, lebanese, bruschetta, tabbouleh, caesar</td>
</tr>
<tr>
<td>豆腐</td>
<td>tofu</td>
<td>boiled, ham, sausage, bacon, kielbasi</td>
</tr>
<tr>
<td>淀粉</td>
<td>starch</td>
<td>flour, beaten, salt, shortening, pwd</td>
</tr>
<tr>
<td>筷子</td>
<td>chopsticks</td>
<td>fork, spatula, toothpick, wooden, knives</td>
</tr>
<tr>
<td>蒸</td>
<td>steam</td>
<td>bake, 350, pans, boil, oblong</td>
</tr>
</tbody>
</table>

English and Chinese recipes, emphasizing that recipe adaptation goes beyond mere translation.

5 Cross-cultural Recipe Adaptation Task

We propose the task of cross-cultural recipe adaptation, which extends the task of machine translation with the requirement of divergence from the source text semantics in order to address cultural differences in the target culture. While translation studies have long considered culture (Bassnett, 2007), this is not yet explored in machine translation. Our matched cross-lingual corpora allow us to inform recipe adaptation by both language and culture simultaneously. In §6 we adopt an end-to-end sequence-to-sequence approach to the task to establish a set of baselines since this is the dominant approach in machine translation.

The evaluation of cultural adaptation should prioritize meaning preservation while allowing divergences in meaning as long as they stem from cross-cultural differences. This subjective criterion is challenging to implement, as cross-cultural differences, and by extension, the task itself, are not well-defined. As common in text generation tasks, we first adopt reference-based automatic evaluation metrics (§5.1). Furthermore, to capture structural similarity between references and predictions, we employ meaning representations for evaluation (§5.2). Crucially, since reference-based metrics are often unreliable for subjective tasks (Reiter, 2018), we additionally perform human evaluation (§5.3).

5.1 Surface-based Automatic Evaluation

We use various metrics to assess the similarity between the generated and reference recipes. We use three overlap-based metrics: BLEU (Papineni et al., 2002), a precision-oriented metric based on token \( n \)-gram overlap and commonly used in machine translation evaluation, ChrF (Popović, 2015), a character-level F-score metric that does not depend on tokenization, and ROUGE-L (Lin, 2004), a recall-oriented metric based on longest common subsequences and widely used in summarization evaluation; and one representation-based metric, BERTScore

\[ \text{For BLEU and ChrF, we use SacreBLEU (Post, 2018) version 2.3.1 with default parameter settings.} \]

\[ \text{For evaluation, we replace newlines with spaces in all reference and generated recipes. We segment Chinese text to words with jieba.} \]
(Zhang et al., 2019b), based on cosine similarity of contextualized token embeddings\textsuperscript{14} and shown to correlate better with human judgments than the above metrics in various tasks.

5.2 Structure-aware Automatic Evaluation

Standard metrics may not effectively capture semantic similarity between texts due to sensitivity to surface form. To address this, we employ graph representations, a favored choice for capturing the flow of cooking actions, tool usage, and ingredient transformations in recipes (Mori et al., 2014; Kiddon et al., 2015; Jermsurawong and Habash, 2015; Yamakata et al., 2016). These allow for an examination of structural differences influenced by language and culture (Wein et al., 2022). Here, we leverage Abstract Meaning Representation (AMR; Banarescu et al., 2013), a general-purpose graph meaning representation, to represent recipes.

To generate AMR graphs, we employ XAMR (Cai et al., 2021),\textsuperscript{15} a state-of-the-art cross-lingual AMR parser that can parse text from five different languages into their corresponding AMR graphs. It is based on a sequence-to-sequence model, utilizing mBART (Liu et al., 2020a) for both encoder and decoder initialization.

To assess the similarity between model-generated and reference texts’ AMRs, we use the Smatch metric (Cai and Knight, 2013), which aligns both graphs and computes the F1 score that measures normalized triple overlap.

5.3 Human Evaluation

While the above automatic metrics provide quantifiable results, they inherently suffer from the limitation of depending on a fixed reference set. In reality, there exist multiple legitimate ways to adapt a recipe. To address this, we propose four criteria for human evaluation, which we conduct on the gold-standard test set.

We have evaluators assess the outputs from all methods, including the human-written adaptations, on four dimensions key to the cultural adaptation of recipes: (1) Grammar—The generated recipe is grammatically sound and fluent; (2) Consistency—The output aligns with the format of a fully executable recipe encompassing coherent title, ingredients, and cooking steps; (3) Preservation—The adapted recipe largely retains the essence of the source recipe, producing a dish akin to the original; (4) Cultural Appropriateness—The generated recipe integrates well with the target cooking culture, aligning with the evaluator’s culinary knowledge and recipe style expectations. Evaluators mark each dimension on a 7-point Likert scale (Likert, 1932), where a higher score indicates superior performance. A single evaluator rates each recipe pair separately and independently.

Crowdsourcing Evaluation. We recruit evaluators on Prolific\textsuperscript{16} and deploy our evaluation platform on the same in-house web application used for human recipe writing (§3.4). To ensure the evaluation validity, we require participants to be native speakers of the target language and proficient in the source language for each adaptation direction. Additionally, participants must successfully undergo a comprehension check, guided by our evaluation tutorial. Each evaluator is required to evaluate two example recipes for the comprehension check and three recipes for our tasks. This rigorous screening process secures the reliability and accuracy of the evaluations conducted for our study.

6 Experiments

Here we describe our recipe adaptation experiments and results, using the CulturalRecipes dataset introduced in §3. Due to their success in machine translation, we experiment with three end-to-end sequence-to-sequence classes of models to adapt recipes across cultures: (finetuned) machine translation models, finetuned multilingual encoder-decoder models, and prompt-based (zero-shot) multilingual language modeling. Additionally, we evaluate the automatic matching approach used in our dataset construction. These will serve as baselines for future work on this task.

6.1 Experimental Setup

We use our silver training set for finetuning in each direction and evaluate on both the silver and gold test sets. We represent a recipe as a concatenation of title, ingredients, and steps, each section prefixed with a heading.

\textsuperscript{14}We rely on bert-base-uncased for representing English text and bert-base-chinese for Chinese text.

\textsuperscript{15}We use the trained AMR parser model from https://github.com/jcyk/XAMR.

\textsuperscript{16}https://www.prolific.co/.
Automatic Matching. Since the source recipes used in the creation of the gold-standard test set are a subsample of the ones found in the silver-standard test set, we have matches for them in the target language retrieved based on title similarity (see §3.3 for a reminder of how the silver-standard test set was constructed). We evaluate these retrieved matches against the gold-standard human-written references, to determine whether title-based retrieval is a viable method for recipe adaptation.

Machine Translation. Recognizing the intrinsic translation component of recipe adaptation between languages, we leverage pre-trained machine translation systems in our experiments. We experiment with opus-mt models (Tiedemann and Thottingal, 2020), which show a strong performance in machine translation. We first evaluate them in zero-shot mode (MT-zs), that is, purely as machine translation models, and additionally after finetuning using our training and validation sets (MT-ft).

Multilingual Language Modeling. We fine-tune multilingual encoder-decoder pre-trained language models on the CulturalRecipes dataset. Such models perform well on translation tasks (Tang et al., 2020) and are generally trained on abundant monolingual as well as parallel data, so they could prove more suitable for the recipe domain and for our ultimate goal, recipe adaptation. We choose mT5-base (Xue et al., 2021), a multilingual multitask text-to-text transformer pre-trained on a Common Crawl-based dataset containing 101 languages, and mBART50 (Tang et al., 2020), a variant of mBART (Liu et al., 2020b) based on a multilingual autoencoder finetuned for machine translation.

Prompting LLMs. Building on the remarkable performance of Multilingual LLMs in zero-shot translation without additional finetuning or in-context learning (Wang et al., 2021), we explore their recipe translation and adaptation capabilities.

We use BLOOM (Scao et al., 2022), an LLM trained on the multilingual ROOTS corpus (Laurenc¸on et al., 2022). Using the ROOTS search tool (Piktus et al., 2023), we find it does not contain our recipe corpora. As BLOOM is an autoregressive language model trained to continue text, we prompt as follows for English→Chinese:

[English recipe] Recipe in English, adapted to an English-speaking audience:

and for Chinese→English:

[Chinese recipe] Recipe in English, adapted to an English-speaking audience:

Further, we experiment with GPT-4 (OpenAI, 2023), and ChatGLM2 (Zeng et al., 2022; Du et al., 2022), state-of-the-art multilingual and Chinese instruction-tuned LLMs (Ouyang et al., 2022). While they have likely been trained on both our recipe corpora (§3.1), they do not benefit from our matching procedure (§3.3) or our newly written human-adapted recipes (§3.4). We prompt them as follows for English→Chinese:

Convert the provided English recipe into a Chinese recipe so that it fits within Chinese cooking culture, is consistent with Chinese cooking knowledge, and meets a Chinese recipe’s style. [English recipe]

and for Chinese→English:

Convert the provided Chinese recipe into an English recipe so that it fits within Western cooking culture, is consistent with Western cooking knowledge, and meets a Western recipe’s style. [Chinese recipe]

Technical Details. For finetuning, we use a batch size of 64 for MT-ft and 32 for mT5-base and mBART50; and a learning rate of 1e-4. We set the maximum sequence length to 512 tokens and finetune models for 30 epochs with early stopping.

---

17We treat these headings as language-invariant meta-text, which is removed in post-processing prior to evaluation.
18Helsinki-NLP/opus-mt-{zh-en/en-zh}.
19google/mt5-base.
20facebook/mbart-large-50.
21bigscience/bloom-7b1, a 7B-parameter model with a 2k-token length limit. Preliminary experiments showed poor results with BLOOMZ-7B, mT0-xxl-mt and FLAN-T5-xxl (Chung et al., 2022), which are finetuned on multitask multilingual prompts (Muennighoff et al., 2022)—they are biased towards short outputs, prevalent in their training tasks.
22gpt-4-0314 via the OpenAI API (8k-token length limit).
23Accessed via FastChat (ChatGLM2-6B).
24Selected among the learning rates {1e-5, 1e-4} for MT-ft, {5e-5, 1e-4} for mT5-base and mBART50; and batch sizes {64, 128} for MT-ft and {32, 64} for mT5-base and mBART50.
after 5 epochs of no improvement in BLEU on the silver validation set. We use two 40GB A100 GPUs for finetuning mT5 and mBART50 and a single one for finetuning MT-ft and for prompting BLOOM. We use the default settings for GPT-4. For ChatGLM2 we set the temperature to 0.7 and the maximum sequence length to 1024 tokens. For generation with all other models, we use a beam of size 3 and a repetition penalty of 1.2; we prevent repeated occurrences of any n-gram of length ≥ 5.

### 6.2 Results

#### Automatic Evaluation on the Silver Test Sets.

As presented in Table 3, we restrict our evaluation on the silver-standard test set to finetuned methods, as a sanity check for their quality under conditions resembling their training setting. We discern that finetuning the MT model considerably improves its performance across all metrics and in both adaptation directions. In Chinese→English, MT-zs emerges as the optimal foundation for finetuning, outperforming the other two methods, mT5, and mBART50, across all metrics. However, English→Chinese displays mixed outcomes, with diverse models excelling in different criteria. Structure-aware automatic evaluation results generally match other automatic results: MT-ft performs best on Chinese→English, while mT5-base performs best on English→Chinese.

---

**Table 3:** Automated evaluation results on the silver test sets using reference-based metrics: SacreBLEU (BLEU), ChrF, R-L (ROUGE), B-Sc (BERTScore)—all token-based, and Smatch—a structure-aware metric assessing AMR graph similarity. Higher scores indicate better performance on all metrics.

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>ChrF</th>
<th>R-L</th>
<th>B-Sc</th>
<th>Smatch</th>
<th># Tok.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese → English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT-zs</td>
<td>6.8</td>
<td>28.7</td>
<td>12.0</td>
<td>54.0</td>
<td>23.7</td>
<td>82.4</td>
</tr>
<tr>
<td>MT-ft</td>
<td><strong>68.9</strong></td>
<td><strong>43.8</strong></td>
<td><strong>22.3</strong></td>
<td><strong>66.4</strong></td>
<td><strong>33.1</strong></td>
<td><strong>98.7</strong></td>
</tr>
<tr>
<td>mT5</td>
<td>60.0</td>
<td>37.2</td>
<td>19.5</td>
<td>62.9</td>
<td>31.0</td>
<td>85.2</td>
</tr>
<tr>
<td>mBART50</td>
<td>44.5</td>
<td>36.0</td>
<td>21.0</td>
<td>63.4</td>
<td>32.1</td>
<td>89.9</td>
</tr>
<tr>
<td>English → Chinese</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT-zs</td>
<td>2.6</td>
<td>9.3</td>
<td>49.7</td>
<td>62.4</td>
<td>20.6</td>
<td>110.6</td>
</tr>
<tr>
<td>MT-ft</td>
<td>38.5</td>
<td>37.1</td>
<td>54.5</td>
<td>71.4</td>
<td>26.8</td>
<td>91.4</td>
</tr>
<tr>
<td>mT5</td>
<td><strong>39.2</strong></td>
<td><strong>36.3</strong></td>
<td><strong>54.9</strong></td>
<td><strong>71.9</strong></td>
<td><strong>27.0</strong></td>
<td><strong>82.1</strong></td>
</tr>
<tr>
<td>mBART50</td>
<td>30.5</td>
<td>32.9</td>
<td><strong>56.2</strong></td>
<td>71.1</td>
<td>25.5</td>
<td>103.2</td>
</tr>
</tbody>
</table>

#### Automatic Evaluation on the Gold Test Sets.

As presented in Table 4, we gain further intriguing insights. The significant performance gap between MT-zs and MT-ft reemphasizes that the recipe pairs in our dataset are not merely translations of each other. Moreover, it underscores the systematic patterns in the matched pairs within our training corpus (reflecting the cultural adaptation of recipes) can indeed be learned via finetuning on retrieved recipes. In this scenario, the LLMs BLOOM, ChatGLM2, and GPT-4 outperform the finetuned methods. Particularly in the Chinese→English direction, LLMs consistently match or surpass the performance of the next best finetuned approach. Notably, a comparison of the average length of model predictions shows a tendency of LLMs to produce longer predictions than their counterparts, with GPT-4 generating double the number of tokens compared to other methods. Interestingly, the retrieval method scores are comparable to the finetuned models in both directions and sometimes even surpass them. Despite this, LLMs continue to prove more effective overall. Smatch scores show performance differences consistent with BERTScore across models for both silver and gold-standard test sets, with the exception that BLOOM slightly outperforms GPT-4 in Chinese→English.

**Table 4:** Automatic reference-based evaluation results on the gold-standard human test sets. † indicates methods without training for the task (zero-shot).

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>ChrF</th>
<th>R-L</th>
<th>B-Sc</th>
<th>Smatch</th>
<th># Tok.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chinese → English</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT-zs†</td>
<td>10.6</td>
<td>6.9</td>
<td>60.8</td>
<td>69.8</td>
<td>29.4</td>
<td>108.0</td>
</tr>
<tr>
<td>MT-ft†</td>
<td>13.6</td>
<td>28.3</td>
<td>53.8</td>
<td>70.5</td>
<td>24.5</td>
<td>88.5</td>
</tr>
<tr>
<td>mT5†</td>
<td>16.6</td>
<td>28.1</td>
<td>53.4</td>
<td>70.7</td>
<td>25.3</td>
<td>78.6</td>
</tr>
<tr>
<td>mBART50†</td>
<td>11.8</td>
<td>25.4</td>
<td>54.8</td>
<td>69.7</td>
<td>23.5</td>
<td>100.3</td>
</tr>
<tr>
<td>BLOOM†</td>
<td>20.0</td>
<td>11.5</td>
<td>50.8</td>
<td>66.4</td>
<td>28.6</td>
<td>154.7</td>
</tr>
<tr>
<td>ChatGLM2†</td>
<td>22.4</td>
<td>11.0</td>
<td>54.3</td>
<td>75.2</td>
<td>28.8</td>
<td>153.2</td>
</tr>
<tr>
<td>GPT-4†</td>
<td>21.1</td>
<td>21.9</td>
<td><strong>61.0</strong></td>
<td>77.8</td>
<td><strong>29.6</strong></td>
<td>213.3</td>
</tr>
<tr>
<td>Retrieval†</td>
<td><strong>32.8</strong></td>
<td><strong>33.6</strong></td>
<td>52.9</td>
<td>68.4</td>
<td>25.0</td>
<td>130.3</td>
</tr>
</tbody>
</table>

---

25We include MT-zs as a reference point to observe the gains from finetuning this model to obtain MT-ft.
Table 5: Human evaluation results on the gold-standard test sets: average and standard deviation across recipes for each method and metric, ranging from 1 to 7. Note that different participants manually adapted (‘Human’) and evaluated the recipes.

<table>
<thead>
<tr>
<th>Method</th>
<th>GRA</th>
<th>CON</th>
<th>PRE</th>
<th>CUL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chinese → English (n = 25)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT-zs</td>
<td>2.6 ±1.5</td>
<td>2.4 ±1.7</td>
<td>2.3 ±1.4</td>
<td>2.7 ±1.6</td>
</tr>
<tr>
<td>MT-ft</td>
<td>4.5 ±1.8</td>
<td>3.7 ±2.0</td>
<td>3.0 ±2.1</td>
<td>4.3 ±2.1</td>
</tr>
<tr>
<td>mT5</td>
<td>4.1 ±2.1</td>
<td>3.8 ±2.1</td>
<td>3.2 ±2.2</td>
<td>3.7 ±2.2</td>
</tr>
<tr>
<td>BLOOM</td>
<td>3.3 ±2.0</td>
<td>3.3 ±2.0</td>
<td>3.4 ±2.0</td>
<td>2.8 ±1.8</td>
</tr>
<tr>
<td>ChatGLM2</td>
<td>4.1 ±2.4</td>
<td>4.3 ±2.2</td>
<td>4.6 ±2.1</td>
<td>4.0 ±2.3</td>
</tr>
<tr>
<td>GPT-4</td>
<td>6.0 ±1.2</td>
<td>6.1 ±1.3</td>
<td>5.9 ±1.0</td>
<td>6.0 ±1.2</td>
</tr>
<tr>
<td>Human</td>
<td>4.2 ±2.1</td>
<td>4.4 ±1.9</td>
<td>4.5 ±1.9</td>
<td>4.6 ±1.9</td>
</tr>
<tr>
<td>Retrieval</td>
<td>5.1 ±1.7</td>
<td>4.9 ±2.0</td>
<td>4.3 ±2.3</td>
<td>3.8 ±2.0</td>
</tr>
</tbody>
</table>

Table 6: Kendall correlation of human evaluation results with automatic metrics. Statistically significant correlations are marked with *, with a confidence level of \( \alpha = 0.05 \) before adjusting for multiple comparisons using the Bonferroni correction (Bonferroni, 1936).

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>ChrF</th>
<th>R-L</th>
<th>B-Sc</th>
<th>Smatch</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chinese → English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MT-zs</td>
<td>0.135</td>
<td>0.250*</td>
<td>0.135</td>
<td>0.257*</td>
<td>0.021</td>
</tr>
<tr>
<td>MT-ft</td>
<td>0.151</td>
<td>0.268*</td>
<td>0.180</td>
<td>0.294*</td>
<td>0.065</td>
</tr>
<tr>
<td>mT5</td>
<td>0.174</td>
<td>0.312*</td>
<td>0.261*</td>
<td>0.260*</td>
<td>0.176</td>
</tr>
<tr>
<td>BLOOM</td>
<td>0.120</td>
<td>0.216*</td>
<td>0.189</td>
<td>0.237*</td>
<td>0.071</td>
</tr>
<tr>
<td>ChatGLM2</td>
<td>0.153</td>
<td>0.255*</td>
<td>0.202*</td>
<td>0.277*</td>
<td>0.079</td>
</tr>
</tbody>
</table>

Human Evaluation. Table 5 showcases the results of human evaluation, with abbreviations GRA, CON, PRE, and CUL representing Grammar, Consistency, Preservation, and Cultural Appropriateness, respectively.\(^{26}\) GPT-4 excels significantly across all metrics in the Chinese→English direction, even surpassing explicit human adaptation. Recipes retrieved from popular websites are a close second in GRA and CON, reflecting their high quality. However, the targeted adaptations written by humans who were explicitly instructed to adapt the source recipe to the target culture, perform better in PRE and CUL. For English→Chinese, GPT-4 remains the top performer only in CUL, while mT5 parallels the retrieved recipes in this metric. Notably, ChatGLM2 surpasses even human writers in CON and PRE, but not in GRA.

Correlation of Automatic Metrics with Humans. To determine the reliability of automatic metrics in assessing the quality of recipe adaptations, we examine their correlation with human evaluations across the four metrics and their average. We use Kendall correlation, which is the official meta-evaluation metric used by WMT22.

As illustrated in Table 6, all cases exhibit a positive correlation, albeit with varying strengths from weak to moderate, and with inconsistent performance between the two adaptation directions. For Chinese→English, ChrF and BERTScore indicate the strongest correlation with the average of all criteria. BERTScore further stands out by demonstrating the highest correlation with each individual criterion. On the other hand, for English→Chinese, BLEU performs comparably well, thus highlighting that the effectiveness of these metrics can vary based on the direction of adaptation. ROUGE-L, however, displays a significantly lower correlation, suggesting its limitations in evaluating recipe adaptations. Finally, we observe that Smatch is not significantly correlated with human judgments, possibly due to noise introduced by parsing errors.\(^{27}\)

CUL presents the weakest correlation with most automatic metrics, underscoring the current limitations of automated evaluations in assessing the quality of recipes.
cultural alignment of recipes, and highlighting the essential role of human evaluators. Notably, correlations for English→Chinese generally exhibit greater strength than Chinese→English. This discrepancy is likely due to the variation in sample sizes between the two directions.

7 Analysis and Discussion

Our findings reinforce previous research asserting the cultural bias of LLMs—specifically GPT-4—towards Western, English-speaking, U.S. culture, as exemplified in the food domain (Cao et al., 2023; Naous et al., 2023; Keleg and Magdy, 2023; Palta and Rudinger, 2023). However, our results also offer a more nuanced perspective. While GPT-4 demonstrates an exceptional ability to adapt to Chinese cuisine, its linguistic and semantic capabilities are outperformed by ChatGLM2 in English→Chinese. To delve deeper into these intriguing results, this section examines the strategies these models employ in the adaptation task.

Quantitative Analysis. Referring back to the analysis from §4, we choose a subset of six words and examine how they are handled by four models (MT-zs, MT-ft, and mT5, and GPT-4). Specifically, we measure the rate of literal translation of these concepts by each model, in the context of the recipes from the silver-standard test set of CulturalRecipes.28 For instance, in adapting from English to Chinese, we identify baking as an English-specific concept. We count the appearances of related terms such as ‘bake’, ‘roast’, ‘broil’, and ‘oven’ in English source recipes, denoted as \( c_{\text{source}} \). For each instance, we tally the occurrences of the direct translation, \( c_{\text{target}} \), from either model predictions or retrieved references. We calculate the literal translation rate as \( \frac{c_{\text{target}}}{c_{\text{source}}} \). Figure 4 visualizes the results for five culturally specific concepts and a universally applicable concept, ‘oil’.

We include ‘oil’ as a sanity check and indeed see that the literal rate of translation is high in both the references and in all model predictions. The references show a low to medium rate of literal translations for the remaining five concepts, confirming their cultural specificity. MT-zs often translates these concepts literally, as could be expected from a machine translation model designed for near-literal translation—the difference is especially noticeable for the concepts ‘steam’ and ‘cheese’. The finetuned models MT-ft and mT5, on the other hand, learn to avoid literal translation, presumably opting for culturally appropriate alternatives instead—for ‘steam’, for example, none of the 12 occurrences of the concept in the source Chinese recipe are literally translated in the predictions of MT-ft and mT5.

An interesting trend emerges in GPT-4 predictions, where literal translations are found at a high rate for all concepts, often close to 100%. While this seems counter-intuitive considering the goal of adapting the culturally specific ingredients and cooking methods, in the next section we find that GPT-4 employs a slightly different strategy than just substituting these ingredients and methods.

Qualitative Analysis. We present a qualitative analysis highlighting the adaptation strategies adopted by models, specifically MT-zs, MT-ft, and GPT-4. The analysis centers on the Chinese recipe shown in Figure 1, with model predictions shown in Table 7. The translation from MT-zs directly incorporates Chinese ingredients not common in English recipes, accompanied by numerous spelling and grammatical errors. The prevalence of errors can be attributed to a dearth of recipe domain representations in the machine translation training data of MT-zs. In contrast, MT-ft offers a notably improved recipe rendition, albeit a wholly different red bean soup from the source recipe. Although this results in minimal content retention, it can be viewed as an extreme cultural adaptation, given the infrequent appearance
Table 7: Case study: English adaptations of the Chinese recipe from Figure 1, with manually highlighted (spelling, grammar or semantic) errors, adaptations to cultural differences, and failures to account for such.

of sweet red bean soup in Western cuisine. However, MT-ft sporadically manifests consistency errors, exemplified in this case by duplicating beans in the ingredient list and parsley in the steps. These findings confirm that the generation of coherent recipes continues to be a challenging endeavor for sequence-to-sequence models, corroborating the findings of prior work (Li et al., 2022).29

Recipe Generation. van Erp et al. (2021) outline potential cross-disciplinary approaches involving NLP and food science, claiming that the analysis of digital recipes is a promising but challenging task. Marin et al. (2019) introduce the Recipe1M dataset (see §3) and Lee et al. (2020) finetune GPT-2 (Radford et al., 2019) on it to create a large English language model, RecipeGPT, capable of generating cooking instructions from titles and ingredients or ingredients from instructions and titles. Majumder et al. (2019) introduce a dataset of 180K English recipes from the website Food.com and a neural model to generate recipes according to user preferences inferred from historical interactions. Contrary to these, we focus on recipe adaptation, where generation is conditioned on a source recipe.

Recipe Adaptation. Donatelli et al. (2021) align recipes for the same dish on the action level using recipe graphs (Yamakata et al., 2016), aiming to adapt recipes to users of different levels of expertise. Morales-Garzón et al. (2021a,b, 2022) propose an unsupervised method to adapt recipes according to dietary preferences by proposing ingredient substitutions using domain-specific word and sentence embeddings. However, they do not modify the recipe steps beyond simple ingredient substitution. Li et al. (2022) build a dataset of 83K automatically-matched recipe pairs for the task of editing recipes to satisfy dietary restrictions. They train a supervised model to perform controlled generation, outperforming RecipeGPT.

8 Related Work

Cultural Adaptation of Text. Cultural adaptation overlaps with style transfer, where the goal

is to change the style of text while preserving the meaning (Jin et al., 2022). In addition to style, cultural adaptation also concerns common ground, values and topics of interest (Hershcovich et al., 2022). Particularly in culture-loaded tasks, it becomes crucial to consider cultural differences (Zhou et al., 2023a,b). While semantic divergences are usually treated as errors in machine translation (Briakou and Carpuat, 2021), cross-cultural translation often requires adaptations that change the meaning, e.g., by adapting entities (Peskov et al., 2021) or by adding explanations (Kementchedjhieva et al., 2020). We share the motivation of this line of work, but for the first time focus on recipes, where cultural adaptation is grounded in clear goals (accessibility to the cook and quality of the resulting dish).
They identify the remaining challenge of “controllable recipe editing using more subtle traits such as cuisines (e.g., making a Chinese version of meatloaf),” which we address here. Antognini et al. (2023), in contrast, propose addressing the same task without paired data, utilizing an unsupervised critiquing module and also outperforming RecipeGPT in both automatic and human evaluation. Liu et al. (2022) present a dataset of 1.5M Chinese recipes and evaluate compositional generalization in neural models in the task of counterfactual generation of recipes with substituted ingredients. They find recipe adaptation to be a challenging task: language models often generate incoherent recipes or fail to satisfy the stated constraints. In contrast, we find that after finetuning pre-trained models on our dataset, the models succeed in the task of cultural adaptation.

9 Conclusion and Future Work

In this work, we studied the task of adapting cooking recipes across cultures. We identified dimensions relevant to this task through a data-driven analysis, including differences in ingredients, tools, methods, and measurement units. We introduced CulturalRecipes, a dataset of paired Chinese and English recipes, and evaluated various adaptation methods. Through our experiments and analysis, we show that models can learn to consider cultural aspects, including style, when adapting recipes across cultures, with some challenges remaining in the level of detail and consistency between the different components of a recipe.

We envision our dataset and baselines will be useful for both downstream applications and further studies of cultural adaptation within and beyond NLP. Automatically adapting recipes from one culture to another could facilitate cross-cultural cross-pollination and broaden the horizons of potential users, serving as a bridge between people through food, and being useful to both novice and experienced cooks. Furthermore, our dataset is a challenging benchmark for language models: besides the complex compositional generalization ability required for recipe adaptation (Liu et al., 2022), it assesses the ability of multilingual language models to adapt to target cultural characteristics, and to construct well-formed and faithful recipes. Lastly, our cross-cultural comparative analysis can be extended to sociological and anthropological research.

Future Work. As acknowledged in §2, the cultural categories we assume are highly simplistic. Future work will expand our datasets to treat finer-grained differences, as well as broaden it to more languages and cultures. It will further investigate the factors that impact recipe adaptation and develop more sophisticated modeling approaches to consider them, beyond the sequence-to-sequence approaches we experimented with here. Finally, our dataset can provide a starting point for related tasks, including recipe classification and retrieval.

Cultural categorization can be a sensitive topic so we have been careful to approach it with respect for the communities involved; we encourage future research in the area to maintain this practice. We hope that our research can contribute to a greater understanding and appreciation of diverse cultural traditions and practices related to food and cooking.

Acknowledgments

The authors extend their sincere gratitude to the reviewers and action editors for their invaluable feedback, which significantly contributed to the improvement of this work. Special thanks are also due to Laura Cabello and Nicolas Garneau for their insightful comments and to Qinghua Zhao and Jingcan Huang for their valuable assistance during our initial human evaluations. We also extend our sincere appreciation to the Department of Food Science, University of Copenhagen, and especially Qian Janice Wang for their contributions as volunteer human review adapters. The authors gratefully acknowledge the HPC RIVR consortium (www.hpc-rivr.si) and EuroHPC JU (eurohpc-ju.europa.eu) for funding this research by providing computing resources of the HPC system Vega at the Institute of Information Science (www.izum.si). Yong Cao and Li Zhou gratefully acknowledge financial support from China Scholarship Council. (CSC No. 202206070002 and No. 202206160052).

References


93


Rensis Likert. 1932. A technique for the measurement of attitudes. Archives of Psychology.


Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary,


Shuo Wang, Zhaopeng Tu, Zhixing Tan, Wenxuan Wang, Maosong Sun, and Yang Liu. 2021. Language models are good translators. https://doi.org/10.48550/ARXIV.2106.13627


