Chinese Idiom Paraphrasing

Jipeng Qiang[1*] **Yang Li**[1] **Chaowei Zhang**[1] **Yun Li**[1]
Yi Zhu[1] **Yunhao Yuan**[1] **Xindong Wu**[2,3]

[1] Yangzhou University, China, [2] Hefei University of Technology, China,
[3] Zhejiang Lab, China
{jpqiang,cwzhang,liyun,zhuyi,yhyuan}@yzu.edu.cn, xwu@hfut.edu.cn

Abstract

Idioms are a kind of idiomatic expression in Chinese, most of which consist of four Chinese characters. Due to the properties of non-compositionality and metaphorical meaning, Chinese idioms are hard to be understood by children and non-native speakers. This study proposes a novel task, denoted as Chinese Idiom Paraphrasing (CIP). CIP aims to rephrase idiom-containing sentences to non-idiomatic ones under the premise of preserving the original sentence's meaning. Since the sentences without idioms are more easily handled by Chinese NLP systems, CIP can be used to pre-process Chinese datasets, thereby facilitating and improving the performance of Chinese NLP tasks, e.g., machine translation systems, Chinese idiom cloze, and Chinese idiom embeddings. In this study, we can treat the CIP task as a special paraphrase generation task. To circumvent difficulties in acquiring annotations, we first establish a large-scale CIP dataset based on human and machine collaboration, which consists of 115,529 sentence pairs. In addition to three sequence-to-sequence methods as the baselines, we further propose a novel infill-based approach based on text infilling. The results show that the proposed method has better performance than the baselines based on the established CIP dataset.

1 Introduction

Idioms, called ''成语'' (ChengYu) in Chinese, are widely used in daily communications and various literary genres. Idioms are a kind of compact Chinese expressions that consist of few words but imply relatively complex social nuances. Moreover, Chinese idioms are often used to describe similar phenomena, events, etc., which means the idioms cannot be interpreted with their literal meanings in some cases. Thus, it has always been a challenge for non-native speakers, and even native speakers, to recognize Chinese idioms (Zheng et al., 2019). For instance, the idiom ''鱼肉百姓'' (YuRouBaiXing) shown in Figure 1, represents ''oppress the people'', instead of its literal meaning - ''fish meat the people''.

In real life, if some people do not understand the meaning of idioms, we have to explain them by converting them into a set of word segments that reflect more intuitive and understandable paraphrasing. In this study, we try to manipulate computational approaches to automatically rephrase idiom-containing sentences into simpler sentences (i.e., non-idiom-containing sentences) for preserving context-based paraphrasing, and then benefit both Chinese-based natural language processing and societal applications.

Since idioms are a kind of obstacles for many NLP tasks, CIP can be used as a pre-processing phase that facilitates and improves the performance of machine translation systems (Ho et al., 2014; Shao et al., 2018), Chinese idiom cloze (Jiang et al., 2018; Zheng et al., 2019), and Chinese idiom embeddings (Tan and Jiang, 2021). Furthermore, CIP-based applications can help specific groups, such as children, non-native speakers, and people with cognitive disabilities, to improve their reading comprehension.

We propose a new task in this study, denoted as Chinese Idiom Paraphrasing (CIP), which aims to rephrase the idiom-containing sentences into fluent, intuitive, and meaning-preserving non-idiom-containing sentences. We can treat the CIP task as a special paraphrase generation task. The general paraphrase generation task aims to rephrase a given sentence to another one that possesses identical semantics but various lexicons or syntax (Kadotani et al., 2021; Lu et al., 2021). Similarly, CIP emphasizes rephrasing the idioms of input sentences to word segments that reflect more intuitive and understandable paraphrasing. In recent decades, many researchers devoted to

*Corresponding author: jpqiang@yzu.edu.cn.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 740–754, 2023. https://doi.org/10.1162/tacl_a_00572
Action Editor: Minlie Huang. Submission batch: 12/2022; Revision batch: 3/2023; Published 7/2023.

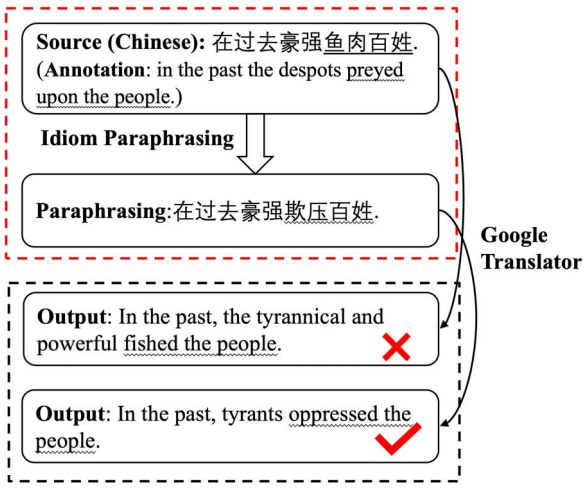

Figure 1: Given a Chinese idiom-containing sentence, we aim to output a fluent, intuitive, and meaning-preserving non-idiom-containing sentence. An idiom-containing sentence is hard to process by NLP applications. For example, this idiom-containing sentence is translated by the newest Google Translator.[1] After processing the idiom-containing sentence using the proposed CIP method, we can output the correct translation. In the example, the idiom is undelined.

paraphrase generation (McKeown, 1979; Meteer and Shaked, 1988) have struggled due to the lack of a reliable supervision dataset (Meng et al., 2021). Inspired by the challenge, we establish a large-scale training dataset in this work for the CIP task.

Contributions. This study produces two main contributions toward the development of CIP systems.

First, a large-scale benchmark is established for the CIP task. The benchmark comprises 115,529 sentence pairs, which of 8,421 are idioms. A recurrent challenge in crowdsourcing NLP-oriented datasets at scale is that human writers frequently utilize repetitive patterns to fabricate examples, leading to a lack of linguistic diversity (Liu et al., 2022). A new large-scale CIP dataset is created in this study by taking advantage of the collaboration between humans and machines.

In detail, we initially divide a large-scale Chinese-English machine translation corpus into two parts (idiom-containing sub-corpus, and non-idiom-containing sub-corpus) by judging if a Chinese sentence contains idioms. Next, we train an English-to-Chinese machine translation (MT) system using the non-idiom-containing sub-corpus. Because the training corpus for the MT system does not include any idioms, the MT system will not translate input English sentences to idiom-containing Chinese sentences. Then, the MT system is deployed to translate English sentences of the idiom-containing sub-corpus to the non-idiom-containing sentences. A large-scale pseudo-parallel CIP dataset can be constructed by pairing the idiom-containing sentences of idiom-containing sub-corpus and the translated non-idiom-containing sentences. Finally, we employ native speakers to validate the generated sentences and modify defective sentences if necessary.

Second, we propose one novel infill-based method to rephrase the input idiom-containing sentence. Since the constructed dataset is used as the training dataset, we treat the CIP task as a paraphrase generation task. We adopt three different sequence-to-sequence (Seq2Seq) methods as baselines: LSTM-based approach, Transformer-based approach, and mT5-based approach, where mT5 is a massively multilingual pre-trained text-to-text Transformer (Xue et al., 2021). Our proposed infill-based method is only required to rephrase the idioms of the sentence, which means that we only need to generate context-based interpretations of idioms, rather than the whole sentence. Specifically, a CIP sentence pair can be processed to produce a (corrupted) input sentence by replacing both the idioms of the source sentence and a corresponding target extracted from the simplified sentence. The mT5-based CIP method is fine-tuned to reconstruct the corresponding target. Experimental results show that, compared with the baselines evaluated on the constructed CIP dataset, our infill-based method can output high-quality paraphrasing of sentences that are grammatically correct and semantically appropriate.

As the use of the Chinese language becomes more widespread, the need for effective Chinese paraphrasing methods may increase, leading to further research and development in this area. The constructed dataset and employed baselines that are used to accelerate this research are open-source, available on Github.[2]

2 Related Work

Paraphrase Generation: Paraphrase generation aims to extract paraphrases of given

[1] translate.google.com. Accessed in: 2022-12-01.

[2] https://www.github.com/jpqiang/Chinese-Idiom-Paraphrasing.

sentences. The extracted paraphrases can preserve the original meaning of the sentence, but are assembled with different words or syntactic structures (McKeown, 1979; Meteer and Shaked, 1988; Zhou and Bhat, 2021).

Most recent neural paraphrase generation methods primarily take advantage of the sequence-to-sequence framework, which can achieve considerable performance improvements compared with traditional approaches (Zhou and Bhat, 2021). Some approaches use reinforcement learning or multi-task learning to improve the quality and diversity of generated paraphrases (Xie et al., 2022). A long-standing issue embraced in paraphrase generation studies is the lack of reliable supervised datasets. The issue can be avoided by constructing manually annotated paired-paraphrase datasets (Kadotani et al., 2021) or designing unsupervised paraphrase generation methods (Meng et al., 2021).

Differ from existing paraphrase generation research, we take our attention to Chinese idiom paraphrasing that rephrases idiom-containing sentences to non-idiom-containing ones.

Text Infilling: Originating from cloze tests (Taylor, 1953), text infilling aims to fill in missing blanks in a sentence or paragraph by making use of the preceding and subsequent text, to make the text complete and meaningful.

Current text infilling methods may be categorized into four groups. GAN-based methods train GANs to ensure that the generator generates highly dependable infilling content that can trick the discriminator (Fedus et al., 2018). Intricate inference-based methods use dynamic programming or gradient search to locate infilling content that is highly probable within its surrounding context (Zaidi et al., 2020). Masked LM-based methods generate infilling content based on its bidirectional contextual word embedding (Shen et al., 2020). LM-based methods fine-tune off-the-shelf LMs in an auto-regressive manner, and some approaches modify the input format by putting an infilling answer after the masked input (Donahue et al., 2020), whereas others do not modify the input format (Zhu et al., 2019). In contrast to the aforementioned methods, our goal in this paper is not only to make the text complete, but also to maintain the sentence's meaning when creating paraphrases. As a result, we employ a sequence-to-sequence framework to identify infilling content.

Idioms: Idiom is an interesting linguistic phenomenon in the Chinese language. Compared with other types of words, most idioms are unique in perspective of non-compositionality and metaphorical meaning. Idiom understanding plays an important role in the research area of Chinese language understanding. Many types of research related to Chinese idiom understanding have been proposed that can benefit a variety of related down-streaming tasks. For example, Shao et al. (2018) focused on evaluating the quality of idiom translation of machine translation systems. Zheng et al. (2019) provided a benchmark to assess the abilities of multiple models on Chinese idiom-based cloze tests, and evaluated how well the models can comprehend Chinese idiom-containing texts. Liu et al. (2019) studied how to improve essay writing skills by recommending Chinese idioms. Tan and Jiang (2021) investigated the tasks on learning and quality evaluation of Chinese idiom embeddings. In this paper, we study a novel CIP task that is different from the above tasks. Since the proposed CIP method can rephrase idiom-containing sentences to non-idiom-containing ones, it is expected that CIP can benefit tasks related to idiom representation and idiom translation.

Pershina et al. (2015) studied a new task of English idiom paraphrases aiming to determine whether two idioms have alike or similar meanings. They collected idioms' definitions in a dictionary and utilized word embedding modelings to represent idioms to calculate the similarity between two idioms. Qiang et al. (2021) proposed a Chinese lexical simplification method, which focuses on replacing complex words in given sentences with simpler and meaning-equivalent alternatives. It is noteworthy that the substitutes in Chinese lexical simplification are all made up of a single word, but an idiom typically cannot be substituted by a single word to express original concepts or ideas.

3 Human and Machine Collaborative Dataset Construction

This section describes the process of constructing a large-scale parallel dataset for CIP. A qualified CIP dataset needs to meet the following two requirements: (1) The two sentences in a sentence pair have to convey the same meaning; and (2) A sentence pair has to contain an idiom-containing

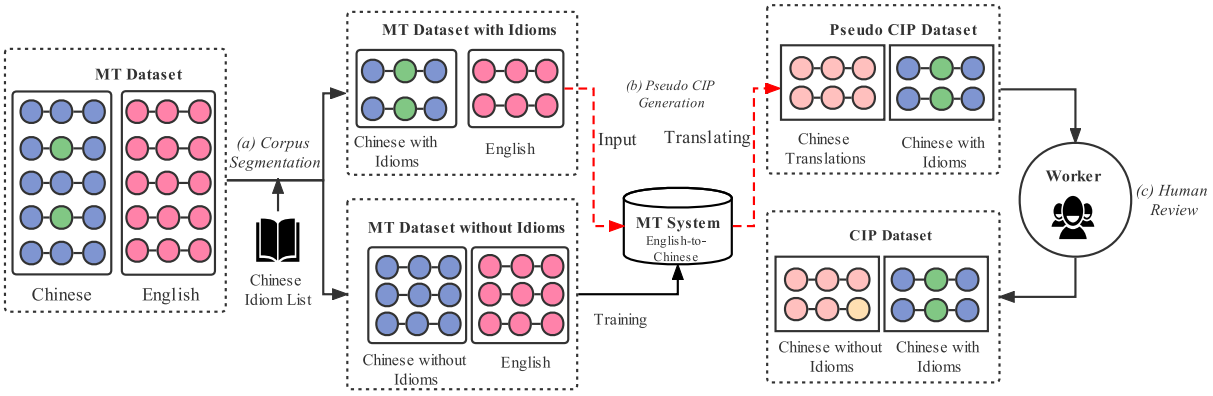

Figure 2: A pipelined illustration of creating a CIP dataset based on a Chinese-English MT corpus. (a) The corpus is split into an idiom-containing sub-corpus and a non-idiom-containing sub-corpus based on a Chinese idiom list. (b) We train a MT system using the non-idiom-containing sub-corpus, and create a pseudo-CIP Dataset by pairing the original Chinese idiom-containing sentences and the translated non-idiom-containing sentences using the trained MT system. (c) We ask human annotators to revise the translated Chinese sentence of the pairs to strengthen the quality of the created CIP dataset.

sentence and an idiom-containing one. We outline a three-stage pipeline for dataset construction, which takes advantage of both the generative strength of machine translation (MT) methods and the evaluative strength of human annotators. Human annotators are generally reliable in correcting examples, but it is challenging while crafting diverse and creative examples at scale. Therefore, we deploy a machine translator to automatically create an initial CIP dataset, and then inquire annotators to proofread each generated instance.

3.1 Pipeline

Figure 2 exhibits the details of the pipeline. Our pipeline starts with an existing English-Chinese machine translation dataset denoted as $\mathcal{D}$. Firstly, we refer to a collect Chinese idiom list $\mathcal{I}$ to split the MT dataset $\mathcal{D}$ into two parts: non-idiom-containing sub-dataset $\mathcal{D}1$ and idiom-containing sub-dataset $\mathcal{D}2$ (Stage 1). All the data items in both $\mathcal{D}1$ and $\mathcal{D}2$ are in forms of sentence pairs. Then, we train a neural machine translation system $\mathcal{M}$ using $\mathcal{D}1$, which can translate English sentences to non-idiom-containing Chinese sentences. Subsequently, we input English sentences in $\mathcal{D}2$ to $\mathcal{M}$ to output non-idiom-containing Chinese sentences. Afterward, the Chinese sentences in $\mathcal{D}2$ and the generated sentences are paired to construct a large-scale initial parallel CIP dataset (Stage 2). Finally, the constructed dataset is reviewed and revised by annotators for quality assurance (Stage 3).

Stage 1: Corpus Segmentation. The English-Chinese MT dataset $\mathcal{D}$ we applied in the research are grabbed from WMT18 (Bojar et al., 2018), which contains 24,752,392 sentence pairs. We extract a Chinese idiom list $\mathcal{I}$ that embraces 31,114 idioms.[3] Since the list enables determining whether the Chinese sentence in a pair contains idioms, $\mathcal{D}$ can be split as $\mathcal{D}1$ and $\mathcal{D}2$. The sub-dataset $\mathcal{D}1$ is used to train a special MT system $\mathcal{M}$ that can translate English sentences to non-idiom-containing Chinese sentences. In our experiments, only 0.2% of the translated Chinese sentences contain idioms (see Table 6). After removing redundant Chinese sentences, the number of sentence pairs in $\mathcal{D}2$ is 105,559.

Stage 2: Pseudo-CIP Dataset. Giving a sentence pair $(\mathbf{c}_i, \mathbf{e}_i)$ in $\mathcal{D}2$, we input the English sentence $\mathbf{e}_i$ into MT system $\mathcal{M}$, and output a Chinese translation $\mathbf{t}_i$. We pair Chinese sentence $\mathbf{c}_i$ and Chinese translation $\mathbf{t}_i$ as a pseudo-CIP sentence pair. Thus, a CIP dataset can be built up by pairing original Chinese sentences and corresponding translated English-to-Chinese ones in $\mathcal{D}2$. The pseudo-CIP dataset $\mathcal{D}2'$ can meet the two requirements of CIP dataset construction. On one hand, the pseudo-CIP data is from the MT dataset, which can guarantee that the paired sentences deliver the same meanings. On another hand, all original sentences include one or more idioms, and all the translated sentences do not contain idioms.

[3] https://github.com/pwxcoo/chinese-xinhua.

		In-domain			Out-of-domain		
		Train	**Dev**	**Test**	**Dev**	**Test**	**Total**
sentence pairs		95,560	5,000	4,999	4,994	4,976	115,529
Source sentence	**token**	3,390,179	173,001	169,793	225,850	221,673	4,180,496
	Avg. sentence length	35	35	34	45	45	36
	All Idioms	102,997	5,423	5,494	5,808	5,800	251,055
	Unique Idioms	7,609	5,225	5,279	5,149	5,128	8,421
Reference sentence	**tokens**	3,454,127	175,083	172,028	239,578	224,907	4,265,723
	Avg. sentence length	36	35	34	48	45	36
Avg. edit disatance		7.85	7.26	7.37	6.21	5.36	7.62

Table 1: The statistics of the CIP dataset.

Freq. Interval		0	[1,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,68)
In	Valid	415	1,787	941	1,159	1,026	67	28
	Test	421	1,814	946	1,171	1,057	60	25
Out	Valid	279	1,871	808	1,120	1,643	62	25
	Test	284	1,854	810	1,108	1,657	66	21

Table 2: Frequency statistics for idiomatic usage in **Dev** and **Test**.

Stage 3: Human Review. As the final stage of the pipeline, we recruit five human annotators to review each sentence pair $(\mathbf{c}_i, \mathbf{t}_i)$ in the pseudo-CIP dataset $\mathcal{D}2'$. These annotators are all undergraduate native Chinese speakers. Given $(\mathbf{c}_i, \mathbf{t}_i)$, annotators are asked to revise and improve the quality of $\mathbf{t}_i$. $\mathbf{t}_i$ is required to be non-idiom-containing and fully meaning-preserving.

3.2 Corpus Statistics

The statistical details of the CIP dataset are shown in Table 1. The dataset $\mathcal{D}2'$ is treated as in-domain data, which contains 105,559 instances including 8,261 different idioms. $\mathcal{D}2'$ is partitioned into three parts: a training set **Train**, a development set **Dev**, and a test set **Test**. The number of instances in **Train**, **Dev**, and **Test** are 95,560, 5,000, and 4,999, respectively.

We observe that both the **Train** and **Test** datasets come from the same distribution. However, when models are deployed in real-world applications, the inference might be performed on the data from different distributions, i.e., out-of-domain (Desai and Durrett, 2020). Therefore, we additionally collected 9,970 sentences with idioms from modern vernacular classics, including prose and fiction, as out-of-domain data, to assess the generalization ability of CIP methods. Unlike the MT corpus, these sentences have no English sentences as their references, we manually modify them to non-idiom-containing sentences with the help of Chinese native speakers.

There are three significant differences between in-domain and out-of-domain data. First, the average length of sentences in in-domain data is around 35 words, while it is about 45 words for out-of-domain data. Second, the average number of idioms in in-domain data is 1.07, which is lower than that of out-of-domain data (i.e., 1.17). Third, the sentence pairs in out-of-domain data need fewer modifications than that in in-domain data. In this case, a lack of linguistic diversity might be taken place due to human annotators often relying on repetitive patterns to generate sentences.

To verify the scalability and generalization ability of the CIP methods, we adopt the following strategy to construct **Dev** and **Test**. We counted the frequency of each idiom in the corpus, where the minimum and the maximum idiom frequency are 1 and 68, respectively. Based on the number of idioms in each frequency interval, we extract the instances into the **Dev** and **Test**. The idiom frequency statistics on the **Dev** and **Test** are shown in Table 2. We can see that those low-frequency idioms occupy a higher proportion of all the idiom occurrences (62.76% and 62.71%

c	约翰并不有钱,他住在海边,深居简出。
e	John is not rich, he live in a simple way near the coast
t	约翰并不有钱,他住在海边,很少出门。
c	她除了演唱外,其余时间则人深居简出。
e	she seldom goes out at other times, except when she sings.
t	她除了演唱外,其他时候很少出门。
c	他们崇尚非暴力、深居简出和远离现代社会的生活方式
e	they believe in nonviolence, simple living and little contact with the modern world.
t	他们信仰非暴力、简单的生活,和远离现代社会的生活方式。
c	绝大多数的时候她都是深居简出，偶尔在公众场合出现。
e	the majority of the time she lives a secluded life, only going out occasionally.
t	她大部分时间过着隐居的生活,偶尔在公众场合出现。
c	我现在爱过幽静的，节俭的深居简出的生活。
e	I chose now to live retired, frugal, and within ourselves.
t	我现在爱过幽静的,节俭的退休的生活。

Table 3: The examples contain the idiom ‘‘深居简出’’ in CIP dataset. **c** and **e** are a machine learning sentence pair, **t** is the CIP reference sentence of **c** generated by collaborating machine translation and human intervention. The words in underlined are idioms, and their translations and their interpretations are marked in wave line.

for low-frequency interval [0,20) in in-domain **Dev** and **Test**). There are 421 and 415 instances containing idioms in in-domain **Dev** and **Test** that are never seen in the **Train**.

3.3 Some Examples in the CIP Dataset

We present some examples of the idiom ‘‘深居简出’’ (reclusive) in the CIP dataset, shown in Table 3. The idiom ‘‘深居简出’’ can be rephrased with different descriptions, displaying the linguistic diversity.

4 Methods

Based on our constructing CIP dataset, the CIP task can be treated as a sentence paraphrasing task (Section 4.1). Additionally, we propose a novel infill-based method to solve it (Section 4.2).

4.1 Paraphrasing for CIP

This can be defined as follows. Given a source sentence $\mathbf{c} = \{c_1, \ldots, c_j, \ldots, c_m\}$ with one or more idioms, we intend to produce a paraphrase sentence $\mathbf{t} = \{t_1, \ldots, t_i, ...t_n\}$. More specifically, **t** is expected to be non-idiom-containing and meaning-preserving, where c_j or t_i refers to a Chinese character. In this study, we suppose to design a supervised method to approach this monolingual machine translation task. We adopt a Sequence-to-Sequence (Seq2Seq) framework that directly predicts the probability of the character-sequential translation from source sentences to target ones (Bahdanau et al., 2015), where the probability is calculated using the following equation 1:

$$P(\mathbf{t} \mid \mathbf{c}) = \prod_{i=1}^{n} P(t_i \mid t_{<i}, \mathbf{c}) \quad (1)$$

where $t_{<i} = t_1, \ldots, t_{i-1}$.

In Section 5, we provide the implementation details of three Seq2Seq methods to handle CIP tasks that are LSTM-based, Transformer-based (Vaswani et al., 2017), and mT5-based (Xue et al., 2021) .

4.2 Infill-based CIP Method

Given a sentence pair {**c**,**t**}, we completely generate the whole target sentence **t** from the original sentence **c** using the Seq2Seq methods. However, CIP merely requires us to rephrase the idioms of the sentence, which means we only expect to generate context-based interpretations of idioms, rather than the whole sentence. Text infilling (Zhu et al., 2019; Xiao et al., 2022) is a task that fills missing text segments of a sentence by a model trained on a large amount of data in a fill-in-the-blank format. Inspired by the work on

text infilling, we propose a novel CIP method, denoted as the infill-based CIP method. Suppose $\bar{\mathbf{c}}$ is an edited sentence of **c** by replacing one idiom into the blank, and the interpretation **y** of the idiom is a sequence of words. The infill-based CIP method aims to generate **y** for filling the blank in $\bar{\mathbf{c}}$.

Extracting Interpretation. Considering that sentence **t** does not only rephrase the part of the idioms, we cannot directly extract the interpretation for each idiom from **t**. We adopt the following method to extract the interpretations of the idioms for {**c**,**t**}.

The interpretation of a given sentence pair {**c**,**t**} is extracted by computing the edit operations. Suppose that **c** is ‘‘约翰踢了我一脚，所以我以牙还牙’’ (John kicked me, so I tit for tat.), and **t** is ‘‘汤姆踢我一脚吧所以我也踢了他一脚.’’ (Tom kicked me, so I kicked him too.), where the characters of the sentence are split by spaces. We construct an edit sequence **Diff** by matching all the words in both **c** and **t** using three edit operations (‘=’,‘−’,‘+’), where ‘=’,‘−’,‘+’ represent ‘keep’, ‘delete’ and ‘add’ operations, respectively.[4] The output **Diff** is ‘‘(‘−’,‘约翰’), (‘+’,‘汤姆’), (‘=’,‘踢’), (‘−’,‘了’), (‘=’, ‘我一脚’), (‘+’, ‘吧’), (‘=’, ’, ‘,所以我’), (‘−’, ‘以牙还牙’), (‘+’, ‘也踢了他一脚’), (‘=’, ‘.’)’’. Specifically, the tuple (‘-’, ‘约翰’) indicates that ‘约翰’ appears in **c** but not in **t**; (‘+’, ‘汤姆’) denotes that ‘汤姆’ appears in **t** but not in **c**; and (‘=’, ‘踢’) represents that ‘踢’ is in both **c** and **t**.

We traverse the edit sequences **Diff** using a rule <‘−’,‘+’>, where the rule means ‘+’ follows by ‘−’ in **Diff**. We get the following two matching sequence pairs <‘约翰’,‘汤姆’> and <‘以牙还牙’,‘也踢了他一脚’>. A sequence pair will be ignored by the model if no idiom is included. In this example, we obtain the sequence pair <‘以牙还牙’,‘也踢了他一脚’>, where ‘‘以牙还牙’’ and ‘‘也踢了他一脚’’ represent an idiom and corresponding interpretation.

Training. Given one sentence pair {**c**,**t**}, we first construct a new sentence pair {$\bar{\mathbf{c}}$ <sep> **c**, **y**}, as shown in Figure 3. Then, we employ the Seq2Seq methods to accomplish this task, as shown in Figure 4. Here, we make two modifications. (1)

[4] https://github.com/paulgb/simplediff.

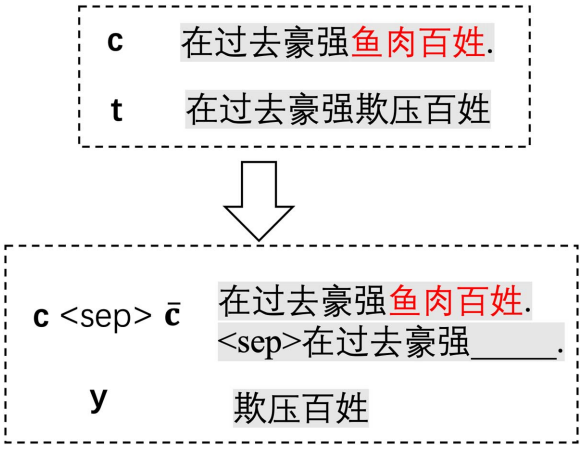

Figure 3: Infill-based CIP method as the sequence-to-sequence task. An sentence pair {**c**,**t**} is transferred into the input ‘‘$\bar{\mathbf{c}}$ <sep> **c**’’ and ‘‘**y**’’.

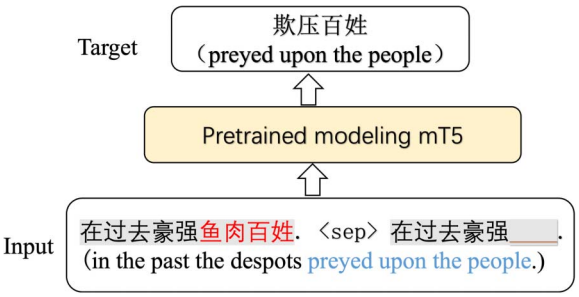

Figure 4: An example of the infill-based CIP method. The input sequence fed into mT5-based Seq2Seq modeling is composed of an original sentence and a target sentence, in which the idiom of the original sentence is replaced by one blank. The interpretation of the idiom is treated as the reference sentence, rather than the target sentence.

If $\bar{\mathbf{c}}$ is directly fed to the encoder, the information of the idiom of **c** is ignored. We concatenate the original **c** and the sentence $\bar{\mathbf{c}}$ as the input sequence. (2) When a sentence has two or more idioms, we construct one {$\bar{\mathbf{c}}$ <sep> **c**, **y**} for each idiom in **c**. We only use one blank for one idiom instead of multiple blanks for all idioms, because we can preserve enough information when generating the sequences to infill the blank.

Given training data consisting of { **c**, $\bar{\mathbf{c}}$, **y**}, it is straightforward to optimize the following objective function according to maximum likelihood estimations:

$$Loss = -\sum_{<\mathbf{c},\bar{\mathbf{c}},\mathbf{y}>\in \mathbf{Train}} \log P(\mathbf{y}|\mathbf{c},\bar{\mathbf{c}};\theta) \quad (2)$$

mT5 is a pre-trained span masked language modeling (MLM) to build a Seq2Seq modeling. In contrast to MLM in BERT (Devlin et al., 2019),

span MLM reconstructs consecutive spans of input tokens and masks them with a blank. With the help of mT5, the proposed method enables reconstructing the idiom I's interpretation of sentence **c** via replacing the idiom with the blank. Therefore, our infill-based method adopts mT5 to learn how to fill the black.

During inference, if a sentence has multiple idioms, we iteratively decode each idiom to a corresponding representation.

Relation to Previous Work. Compared with the sentence paraphrasing task (Zhou and Bhat, 2021; Xie et al., 2022), our infill-based method only requires us to rephrase the idioms of the sentence, rather than the whole sentence. Actually, our method is inspired by the text infilling method of Zhu et al. (2019). But our method is different from the existing text infilling method, because our aim is to rephrase the original sentence, and the aim of text infilling is to make the text complete and meaningful.

5 Experiments

5.1 Experiment Setup

Implementation Details. In this experiment design, four CIP methods are deployed, including: LSTM-based Seq2Seq modeling (LSTM), Transformer-based Seq2Seq modeling (Transformer), mT5-based Seq2Seq modeling (mT5), and infill-based CIP method (Infill). We implement LSTM and Transformer methods using fairseq (Ott et al., 2019). mT5 and Infill methods are mT5-based, and are fulfilled using HuggingFace transformers (Wolf et al., 2020). Furthermore, the sentence tokenization is accomplished using the Jieba Chinese word segmenter[5] and BPE tokenization. The size of the vocabulary is set to 32K. The LSTM-based Seq2Seq method adopts the Adam optimizer configured with $\beta = (0.9, 0.98)$, $3e^{-4}$ learning rate, and 0.2 dropout rate. The Transformer-based Seq2Seq method maintains the hyperparameters of the base Transformer (Vaswani et al., 2017) (base), which contains a six-layered encoder and a six-layered decoder. The three parameters (β of Adam optimizer, learning rate, and dropout rate) in the Transformer-based method are equivalent to those in the LSTM-based method. It's noteworthy that the learning rate is gradually increased to $3e^{-4}$ by 4k steps and correspondingly decays according to the inverse square root schedule. For mT5 and Infill, we adopt the mt5 version that is re-trained on Chinese corpus.[6] We train the three methods with the Adam optimizer (Kingma and Ba, 2015) and an initial learning rate of $3e^{-4}$ up to 20 epochs using early stopping on development data. The training will be stopped when the accuracy on the development set does not improve within 5 epochs. We used a beam search with 5 beams for inference.

Metrics. As we mentioned above, the CIP task can be treated as a sentence paraphrasing task. Therefore, We apply four metrics to evaluate sentence paraphrasing task namely, BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2020), and ROUGE-1 and ROUGE-2 (Lin, 2004). BLEU is a widely used machine translation metric, which measures opposed references to evaluate lexical overlaps with human intervention (Papineni et al., 2002). BERTScore is chosen as another metric due to its high correlation with human judgments (Zhang et al., 2020). Compared to BLEU, BERTScore is measured using token-wise cosine similarity between representations produced by BERT. We measure semantic overlaps between generated sentences and reference ones using ROUGE scores (Lin, 2004). ROUGE is often used to evaluate text summarization. The two metrics ROUGE1 and ROUGE2 refer to the overlaps of unigram and bigram between the system and reference summaries, respectively.

Since generated paraphrases often only need to rewrite the idioms of the original sentence, evaluating the whole sentence cannot accurately reflect the quality of paraphrase generation. In order to better evaluate the quality of idiom paraphrasing, we only evaluate the rewrite part of the generating paraphrases instead of the whole sentence using the above metrics. Specifically, given an original sentence **c**, a reference sentence **t**, and the generated paraphrase sentence **u**, we first find the common words in all the sentences **c**, **t**, and **u**, and remove the common words from **t** and **u**. We evaluate the remaining words of **t** and **u** using the above metrics, denoted as (BERT-E, BERTScore-E, ROUGE1-E, and ROUGE2-E).

Baselines. In this research, we adopt three Seq2Seq methods to handle CIP tasks that are

[5] https://github.com/fxsjy/jieba.

[6] www.github.com/ZhuiyiTechnology/t5-pegasus.

Method	BLEU/BLEU-E	BERTS/BERTS-E	ROU1/ROU1-E	ROU2/ROU2-E
Re-translation	27.37/2.67	78.73/64.43	57.01/24.46	31.93/5.32
BERT	74.75/1.39	91.53/62.99	83.05/18.36	73.64/5.11
LSTM	81.99/31.87	93.79/78.73	87.83/55.52	80.20/43.40
Transformer	82.19/32.58	94.00/79.41	88.16/56.70	80.50/44.58
mT5	82.98/33.87	94.22/80.36	88.13/57.44	80.78/45.89
Infill	**83.55/(34.26 ± 0.02)**	**94.46/(80.94±0.05)**	**88.68/(58.44±0.03)**	**81.57/(47.96±0.03)**

Table 4: The results of different methods on the in-domain test set using the metrics: BLEU, BERTS, ROUGE1, and ROUGE2, where BERTS, ROU1, and ROU2 refer to BERTScore, ROUGE-1, and ROUGE-2, respectively. ''±'' means the standard deviation of five runs.

LSTM-based, Transformer-based, and mT5-based models, respectively. We additionally provide two zero-shot methods that can facilitate solving the CIP problem, namely, Re-translation and BERT-CLS.

(1) The LSTM-based Seq2Seq method is a basic Seq2Seq method, which uses an LSTM (long short-term memory [Hochreiter and Schmidhuber, 1997]) to convert a sentence to a dense, fixed-length vector representation. In contrast to vanilla form of RNNs, LSTM can handle long sequences, but it fails to maintain the global information of the sequences.

(2) The Transformer-based Seq2Seq method (Vaswani et al., 2017) is a state-of-the-art Seq2Seq method that has been widely adopted to process various NLP tasks, such as machine translation, abstractive summarization, etc. Transformer applies a self-attention mechanism that directly models the relationships among all words of an input sequence regardless of words' positions. Unlike LSTM, Transformer handles the entire input sequence at once, rather than iterating words one by one.

(3) mT5 is a Seq2Seq method that uses the framework of Transformer. Currently, most downstream NLP tasks build their models by fine-tuning pre-trained language models (Raffel et al., 2020). mT5 is a massively multilingual pre-trained language model that is implemented in a form of unified "text-to-text" to process different downstream NLP problems. In this study, we fine-tune the mT5-based approach to handle CIP task.

(4) Re-translation is implemented by utilizing the back-translation techniques of machine translation methods. We first translate an idiom-containing sentence to an English sentence using an efficient Chinese-English translation system, and then translate the generated English sentence using our trained English-Chinese translation system (introduced by Section 3.1) to generate a non-idiom-containing Chinese sentence. The Chinese-English translation system can be easily accessed online.[7] The trained English-Chinese translation system is a transformer-based Seq2Seq method.

(5) BERT-CLS is an existing BERT-based Chinese lexical simplification method (Qiang et al., 2021). In this task, an idiom is treated as a complex word that will be replaced with a simpler word.

5.2 Performance of CIP Methods

Table 4 summarizes the evaluation results on our established CIP dataset using two types of metrics. The supervised CIP methods (LSTM, Transformer, mT5, and Infill) are significantly better than two zero-shot methods (Re-translation and BERT) in perspectives of the four metrics. The results reveal that the dataset is a high-quality corpus, which can help to benefit CIP task.

Table 4 and Table 5 show that the performance of LSTM-based baseline is inferior to the other three baselines on in-domain and out-of-domain test datasets. In general, the two mT5-based CIP methods (mT5 and Infill) outperform the other two methods (LSTM and Transformer), which suggests that CIP methods fine-tuned on mT5 can improve CIP performance. It is observed that Infill yields the best results on the in-domain test set compared with other CIP methods, which verifies that Infill is quite effective. On the out-of-domain test set, BERT-based method achieves the best results on ROU1 and ROU2, and obtains the worst results on ROU1-E and ROU2-E, because

[7] huggingface.co/Helsinki-NLP/opus-mt-zh-en.

Method	BLEU/BLEU-E	BERTS/BERTS-E	ROU1/ROU1-E	ROU2/ROU2-E
Re-translation	13.65/0.87	72.63/61.87	47.18/22.73	19.47/2.65
BERT	84.95/1.63	94.79/63.07	**89.84**/19.95	**85.10**/5.99
LSTM	81.20/7.81	93.50/67.03	87.66/31.08	81.69/14.14
Transformer	80.14/8.07	93.55/67.36	87.63/31.62	81.56/14.48
mT5	84.76/9.29	94.58/68.05	89.20/**31.82**	83.76/14.64
Infill	**86.60/(10.68±0.03)**	**94.98/(68.54±0.05)**	89.70/(31.10±0.03)	84.89/(**15.85**±0.02)

Table 5: The results of different methods on the out-of-domain test set. ''±'' means the standard deviation of five runs.

Method	In				Out			
	Simp.	Meaning	Fluency	Avg	Simp.	Meaning	Fluency	Avg
Reference	4.41	4.26	4.29	4.32	4.00	3.86	3.77	3.88
BERT	3.28	2.24	2.58	2.70	2.98	2.06	2.26	2.44
LSTM	3.58	3.33	3.27	3.39	3.26	3.02	2.89	3.06
Transformer	3.48	3.29	3.22	3.33	3.16	3.01	2.89	3.02
mT5	3.85	3.57	3.62	3.68	3.23	3.02	2.97	3.07
Infill	**4.02**	**3.78**	**3.81**	**3.87**	**3.72**	**3.49**	**3.42**	**3.54**

Table 6: The results of human evaluation. ''Simp.'' denotes ''simplicity'', ''Avg'' denotes ''average''.

it makes minor modifications on the source sentence. It means that the type of metrics (BLEU-E, BERTS-E, ROU1-E, and ROU2-E) are more reasonable as the evaluation metrics for CIP task. Our proposed Infill-based method is still the best option for CIP task on the out-of-domain test set.

Our proposed method Infill is superior to the baselines in several key ways. First, our approach is more efficient, allowing us to achieve better results in less time, because infill-based method only needs to rephrase the idioms of the input sentence. Second, our method is more robust, since it achieves the best results in both in-domain and out-of-domain test sets. Finally, our method has been extensively tested and validated, giving us confidence in its reliability and accuracy through human evaluation. Overall, our method represents a significant improvement over the existing baselines and is the best option for solving the CIP problem at hand.

5.3 Human Evaluation

For further evaluating the CIP methods, we adopt human evaluation to analyze the deployed CIP methods. We choose 184 sentences from the in-domain test set and 197 sentences from the out-of-domain test set. To verify the scalability and generalization ability of the CIP methods, we show the performance of CIP methods when they are used to solve the problems where the idioms are never seen in the corpus. Therefore, we choose 49 and 48 test sentences for in-domain and out-of-domain test sets containing idioms that do not appear in the training set, respectively. We ask five native speakers to rate each generated sentence using three features: simplicity, meaning, and fluency. The five-point Likert scale is adopted to rate these features, and the average scores of

Method	**In**	**Out**
Re-translation	99.82%	99.86%
BERT	78.18%	69.21%
LSTM	85.64%	84.65%
Transformer	87.26%	85.13%
mT5	87.54%	72.07%
Infill	87.98%	81.29%

Table 7: The proportion between the times of idiom paraphrasing and the number of all idioms.

Sent1	隋朝开始开科取士，最初亦为取秀才。
English	sui Dynasty began to open branches to take disabilities, also taking the first scholar.
Reference	隋朝开始举行科举考试，最初亦为取秀才。
LSTM, Trans., mT5	隋朝开始开科取士，最初亦为取秀才。
Infill	隋朝开始科举，最初亦为取秀才。
Sent2	他们说我所尝试着去获得的仅仅是黄梁一梦。
English	they said what I tried to attain was a pipe dream and nothing more.
Reference	他们说我所尝试着去获得的仅仅是一个美梦。
LSTM	他们说我所尝试着去获得的仅仅是黄梁一梦。
Trans.	他们说我所尝试着去获得的仅仅是黄梦。
mT5	他们说我所尝试着去获得的仅仅是梦。
Infill	他们说我所尝试着去获得的仅仅是一场梦。
Sent3	你已经不小了，应该能够顶门立户了。
English	You are not young anymore, you should be able to start a family.
Reference	你已经不小了，应该能够当家了。
LSTM, Trans.	你已经不小了，应该能够顶门立户了。
mT5	你已经不小了，应该能够结婚了。
Infill	你已经不小了，应该能够有担当了。

Table 8: The outputting paraphrasing of CIP methods when the idioms do not appear in the training set. ‘‘Trans.’’ denotes ‘‘Transformer’’.

the features are calculated correspondingly. (1) Simplicity is responsible for evaluating whether re-paraphrased idioms of generated sentences are easily understandable, which means idioms in the original sentences should be rewritten with simpler and more common words. (2) Meaning assesses whether generated sentences preserve the meaning of the original sentences. (3) Fluency is used to judge if a generated sentence is fluent and does not contain grammatical errors.

The results of the human evaluation are shown in Table 6. We calculate the scores of annotated sentences **t**, denoted as Reference. We see that the infill-based mT5 method outperforms other methods on in-domain and out-of-domain test sets, which means Infill is an effective method on the CIP task. The conclusions are consistent with the results using automatic metric. Compared with Reference, our method has a significant potential for improvement.

Additionally, we calculated the inter-annotator agreement among different annotators. Specifically, we computed Fleiss' Kappa (Fleiss, 1971) scores for different domain test sets. The scores are 0.199 and 0.097 in in-domain and out-of-domain test sets, respectively. This indicates that the evaluation of Chinese idiom paraphrasing was relatively subjective but still managed to achieve a modest level of agreement. We acknowledge the limitations of human judgment in evaluating the quality of paraphrasing and believe that the diversity of opinions among raters is a valuable insight into the complexity of the CIP task.

5.4 Proportion of Idiom Paraphrasing

CIP aims to rephrase an input idiom-containing sentence to a meaning-preserving and non-idiom-containing sentence. In this subsection, we count the number of idiom-containing sentences that are rephrased to non-idiom-containing sentences. The results are shown in Table 7. The result shows that Re-translation achieves the best results, which can rephrase almost all idioms to non-idiom-containing representations. That means that our idea on CIP dataset construction using machine translation method is feasible. Theoretically, if the trained English-Chinese machine translation method (Stage 2 in the pipeline) can output high-quality results, we do not need to ask annotators to optionally revise Chinese translations. We observe that the proportions of these CIP methods (LSTM, Transformer, mT5, and Infill) are nearly 90%, which means they have great potential for dealing with idiom paraphrasing. Moreover, a

small number of idioms cannot be rephrased, because some idioms are simple, thereby are retained in the training set.

5.5 Case Study

The in-domain test set contains idioms that are never seen in **Train**. We show the paraphrasing results of different CIP methods in Table 8.

Our method consistently outperforms all other approaches in the case study. We found that LSTM-based and Transformer-based methods tend to retain or output part of the idioms, because they cannot learn any knowledge of these idioms from the training corpus. We found that both mT5-based and Infill-based methods based on the pretrained language model mT5 can generate correct interpretations for some of the idioms, as the mT5 model has learned the knowledge of these idioms. The mT5-based method generates a whole new sentence for the original sentence, which can lead to some incorrect interpretations. In contrast, the Infill-based method only rephrases the idioms within the sentence based on their context, which can produce higher-quality interpretations compared to the mT5-based method.

5.6 The Translations of Chinese Idioms

Not only do idioms present a challenge for people to understand, but they also present a greater challenge for Chinese-based NLP applications. Here, we use Chinese-English machine translation as an example of an NLP application to evaluate the usefulness of CIP methods. Given an input sentence containing an idiom, we first use our CIP method as a preprocessing technique to rephrase the sentence, and then translate the paraphrased version into an English sentence.

We give some examples to compare the differences, and the results are shown in Table 9. Because many idioms cannot be translated with their literal meaning, our method helps to identify and paraphrase these idioms, making them easier for the machine translation system to process.

Sent1	她的容貌如花似月。
Tran1	Her face is like a flower like a moon.
Para	她的容貌很美丽。
Tran2	Her face is beautiful.
Sent2	一提起癌，人们便谈虎色变。
Tran1	When it comes to cancer, people talk about it.
Para	一提起癌，人们便吓得脸色就变了。
Tran2	When it comes to cancer, people's faces change with fear.
Sent3	你喜欢趁机混水摸鱼。
Tran1	You like to fish in the water.
Para	你喜欢趁机乘着混乱捞取利益。
Tran2	You like to ride the chaos to your advantage.
Sent4	你这是在钻牛角尖?
Tran1	Are you digging the horns?
Para	你这是在固执地坚持?
Tran2	Are you stubbornly insisting?
Sent5	这是一片钟灵毓秀的土地。
Tran1	This is a land of Zhongling Yuxiu.
Para	这是一片景色优美的土地。
Tran2	This is a beautiful land.
Sent6	如此刁钻的问题简直是强人所难。
Tran1	Such a tricky question is simply beyond the reach of a strong man.
Para	如此刁钻的问题简直是勉强别人。
Tran2	Such tricky questions are simply forcing others.
Sent7	好极了！然而，你应该趁热打铁。
Tran1	great! However, you should strike while the iron is hot.
Para	好极了！然而，你应该抓紧时机。
Tran2	great! However, you should hurry up.
Sent8	我愿为我的好朋友赴汤蹈火。
Tran1	I am willing to go through fire and water for my good friend.
Para	我愿为我的好朋友历经艰险。
Tran2	I am willing to go through hardships for my good friend.

Table 9: The following examples demonstrate how our CIP method can improve a Chinese-English machine translation system. ‘‘Tran1’’ is the translation of the original sentence using Google Translate,[8] ‘‘Para’’ is the paraphrased version generated by our infill-based method, and ‘‘Tran2’’ is the translation of the paraphrased sentence using Google Translate.

6 Conclusion

In this paper, we propose a novel Chinese idiom paraphrasing (CIP) task, which aims to rephrase sentences containing idioms into non-idiomatic versions. The CIP task can be treated as a special case of paraphrase generation and can be addressed using Seq2Seq modeling. We construct

[8] https://translate.google.com/. Accessed in: 2022-12-01.

a large-scale training dataset for CIP by taking the collaborations between humans and machines. Specifically, we first design a framework to construct a pseudo-CIP dataset and then ask workers to revise and evaluate the dataset. In this study, we deploy three Seq2Seq methods and propose one novel CIP methods (Infill) for the CIP task. Experimental results reveal that our proposed methods trained on our dataset can yield good results. This could have a positive impact on the performance of machine translation systems, as well as other natural language processing applications that involve Chinese idioms. In our subsequent research, our proposed methods will be used as strong baselines, and the established dataset will also be used to accelerate the study on this topic.

Acknowledgments

This research is partially supported by the National Natural Science Foundation of China under grants 62076217, 62120106008, and 61906060, and the Blue Project of Yangzhou University.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Ond rej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 272–307, Belgium, Brussels. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6401

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.21

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186. https://https://doi.org/10.18653/v1/N19-1423

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501.

William Fedus, Ian J. Goodfellow, and Andrew M. Dai. 2018. Maskgan: Better text generation via filling in the _____. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 – May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378. https://doi.org/10.1037/h0031619

Wan Yu Ho, Christine Kng, Shan Wang, and Francis Bond. 2014. Identifying idioms in Chinese translations. In *LREC*, pages 716–721.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735, PubMed: 9377276

Zhiying Jiang, Boliang Zhang, Lifu Huang, and Heng Ji. 2018. Chengyu cloze test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–158.

Sora Kadotani, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Edit distance based curriculum learning for paraphrase generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 229–234. https://doi.org/10.18653/v1/2021.acl-srw.24

Diederik P. Kingma, and Jimmy Ba. 2015. A method for stochastic optimization. https://doi.org/10.48550/arXiv.2203.16634

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuanchao Liu, Bo Pang, and Bingquan Liu. 2019. Neural-based Chinese idiom recommendation for enhancing elegance in essay writing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5522–5526. Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1552

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.22

Kathleen R. McKeown. 1979. Paraphrasing using given and new information in a question-answer system. In *17th Annual Meeting of the Association for Computational Linguistics*, pages 67–72, La Jolla, California, USA. Association for Computational Linguistics.

Yuxian Meng, Xiang Ao, Qing He, Xiaofei Sun, Qinghong Han, Fei Wu, Chun Fan, and Jiwei Li. 2021. ConRPG: Paraphrase generation using contexts as regularizer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2551–2562. https://doi.org/10.18653/v1/2021.emnlp-main.199

Marie Meteer and Varda Shaked. 1988. Strategies for effective paraphrasing. In *COLING Budapest 1988 Volume 2: International Conference on Computational Linguistics*, pages 431–436. https://doi.org/10.3115/991719.991724

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53. https://doi.org/10.18653/v1/N19-4009

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. https://doi.org/10.3115/1073083.1073135

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Idiom paraphrases: Seventh heaven vs cloud nine. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 76–82. Lisbon, Portugal. Association for Computational Linguistics. https://doi.org/10.18653/v1/W15-2709

Jipeng Qiang, Xinyu Lu, Yun Li, Yun-Hao Yuan, and Xindong Wu. 2021. Chinese lexical simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1819–1828. https://doi.org/10.1109/TASLP.2021.3078361

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Yutong Shao, Rico Sennrich, Bonnie Webber, and Federico Fancellu. 2018. Evaluating machine translation performance on Chinese idioms with a blacklist method. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Tianxiao Shen, Victor Quach, Regina Barzilay, and Tommi Jaakkola. 2020. Blank language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5186–5198.

Minghuan Tan and Jing Jiang. 2021. Learning and evaluating Chinese idiom embeddings.

In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1387–1396.

Wilson L. Taylor. 1953. ''Cloze procedure'': A new tool for measuring readability. *Journalism Quarterly*, 30(4):415–433. https://doi.org/10.1177/107769905303000401

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yanling Xiao, Lemao Liu, Guoping Huang, Qu Cui, Shujian Huang, Shuming Shi, and Jiajun Chen. 2022. BiTIIMT: A bilingual text-infilling method for interactive machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1958–1969, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.138

Xuhang Xie, Xuesong Lu, and Bei Chen. 2022. Multi-task learning for paraphrase generation with keyword and part-of-speech reconstruction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1234–1243.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Najam Zaidi, Trevor Cohn, and Gholamreza Haffari. 2020. Decoding as dynamic programming for recurrent autoregressive models. In *International Conference on Learning Representations*.

Yuhui Zhang, Chenghao Yang, Zhengping Zhou, and Zhiyuan Liu. 2020. Enhancing transformer with sememe knowledge. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 177–184, Online. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A large-scale Chinese IDiom dataset for cloze test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787. https://doi.org/10.18653/v1/P19-1075

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086. https://doi.org/10.18653/v1/2021.emnlp-main.414

Wanrong Zhu, Zhiting Hu, and Eric P. Xing. 2019. Text infilling. *CoRR*, abs/1901.00158. https://doi.org/10.48550/arXiv.1901.00158