Food for Thought: How can we exploit contextual embeddings in the translation of idiomatic expressions?

Lukas Santing[1], **Ryan Sijstermans**[1], **Giacomo Anerdi**[1],

Pedro Jeuris[1], **Marijn ten Thij**[1] and **Riza Batista-Navarro**[1,2]

[1]Department of Advanced Computing Sciences, Maastricht University, The Netherlands
[2]Department of Computer Science, The University of Manchester, UK

Abstract

Idiomatic expressions (or idioms) are phrases where the meaning of the phrase cannot be determined from the meaning of the individual words in the expression. Translating idioms between languages is therefore a challenging task. Transformer models based on contextual embeddings have advanced the state-of-the-art across many domains in the field of natural language processing. While research using transformers has advanced both idiom detection as well as idiom disambiguation, idiom translation has not seen a similar advancement. In this work, we investigate two approaches to fine-tuning a pretrained Text-to-Text Transfer Transformer (T5) model to perform idiom translation from English to German. The first approach directly translates English idiom-containing sentences to German, while the second is underpinned by idiom paraphrasing, firstly paraphrasing English idiomatic expressions to their simplified English versions before translating them to German. Results of our evaluation show that each of the approaches is able to generate adequate translations.

1 Introduction

In the past decade, we have seen an increase in the accuracy of machine translation (MT) approaches (Wang et al., 2021). Some of the contributing factors to this increase is the introduction of the attention mechanism and contextual embedding models (Liu et al., 2020), as well as the wider availability of datasets. According to Škvorc et al. (2022) however, the same increase in accuracy has not been achieved for idiom translation. Idioms are defined as "a group of words established by usage as having a meaning not deducible from those of the individual words" (University of Oxford, 2022).

Since datasets used for MT are, in general, not rich in idioms (Fadaee et al., 2018; Saxena and Paul, 2020; Zhou et al., 2022; Škvorc et al., 2022; Eryiğit et al., 2022), MT models can suffer from this by not being able to distinguish between an idiom and an expression that can be interpreted literally. This can result in a wrong or meaningless translation, as can be seen in Figure 1. However, there are also idioms which can be interpreted literally, depending on the context in which it was used. For instance, the idiom "breaking the ice" could have an idiomatic meaning "to get a conversation started", but its literal meaning is also valid, e.g., in the context of someone breaking ice cubes for a cocktail. Such idiomatic expressions make it even more challenging for models to correctly translate sentences containing them. To further complicate the issue, some multi-word expressions (MWEs) such as "to pass on" and "to come out", are often used in their idiomatic sense but can also be used in their literal sense.

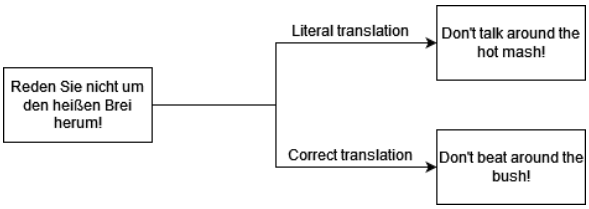


Figure 1: Example of an idiom-containing German sentence with its wrong (literal) and correct translations in English.

With the emergence of the attention mechanism (Yu et al., 2020), transformer models and contextual embeddings (Devlin et al., 2018) came the rapid advancement of the state of the art in many NLP tasks, e.g., question answering and machine translation (Raffel et al., 2020). In this work, we aim to improve the translation of idiomatic expressions by employing contextual embeddings.

We focus on investigating two different approaches for fine-tuning transformer models for the translation of sentences containing idiomatic expressions. Parallel corpora containing idiomatic expressions are scarce (Fadaee et al., 2018; Saxena and Paul, 2020; Zhou et al., 2022; Škvorc et al.,

Proceedings of the 3rd Workshop on Figurative Language Processing (FLP), pages 100 - 110

2022; Eryiğit et al., 2022), but owing to the availability of a parallel corpus of idiom-containing English sentences and their corresponding German translations (Fadaee et al., 2018), we have chosen English and German as our source and target languages, respectively. The first approach utilises this dataset for idiom-to-idiom translation, i.e., translation of an idiom-containing sentence in English to its equivalent idiom-containing sentence in German. The second approach, meanwhile, is based on idiom paraphrasing, i.e., conversion of an English idiom-containing sentence to its paraphrase, followed by translation of the latter to German. On top of assessing the performance of these approaches based on evaluation metrics, we designed a strategy for human-based evaluation to determine: (1) how fluent their translations are in German, and (2) how well their translations preserve the meaning of source sentences.

To the best of our knowledge, ours is the first work to investigate the extent to which transformer models, specifically the Text-to-Text Transfer Transformer (T5) kind (Raffel et al., 2020), can translate idiom-containing sentences from one language to another. Based on the transformer encoder-decoder architecture (Vaswani et al., 2017), T5 provides a unified framework for casting many NLP problems (e.g., text classification, question answering) as a sequence-to-sequence learning task, and thus lends itself well to the problem of translating idiomatic expressions.

2 Related Work

Neural machine translation (NMT) models (Isabelle et al., 2017) and statistical machine translation (SMT) models (Salton et al., 2014a) have shown difficulty in translating idiomatic expressions (Chakrawarti et al., 2017; Dankers et al., 2022). The meaning of an idiom is generally different from the joint meaning of the words composing it, and therefore translation models tend to make errors from the literal translation of individual words.

In recent years, a number of transformer models, e.g., BERT (Devlin et al., 2018), BART (Lewis et al., 2019) and T5 (Raffel et al., 2020), have been successfully applied to a wide range of natural language processing tasks. The advantage of these models is that, for any word (token), they use a contextual embedding representation which is based not only on the word itself, but also on its context (i.e., surrounding words to the left and right). They have achieved ground-breaking results in almost every NLP task (Liu et al., 2020). Context is key for the comprehension of idiomatic expressions, hence such contextual embeddings could potentially be helpful in understanding them. While the use of transformer models to understand idiomatic expressions has been explored in several papers (Kurfali and Östling, 2020; Zhou, 2021; Zhou et al., 2022; Tan and Jiang, 2021; Škvorc et al., 2022), very little research has been done on idiom translation based on these models.

There are multiple tasks involved in the translation of idiom-containing sentences. The first one involves the identification of idiomatic expressions within a sentence (Fazly et al., 2009). Škvorc et al. (2022), for instance, showed that transformer models can be used to successfully identify idiomatic expressions. Idiom identification is followed by sense disambiguation, which involves determining whether an idiom is used literally or idiomatically in the containing sentence (Sporleder and Li, 2009; Kurfali and Östling, 2020; Tan and Jiang, 2021). Transformers have also advanced the state of the art in this task (Kurfali and Östling, 2020; Tan and Jiang, 2021). A further task is the translation or paraphrasing of idioms, depending on the intended application. Much research has been shown to attempt paraphrasing idioms to replace them with their literal meaning (Liu and Hwa, 2016; Zhou, 2021; Zhou et al., 2022; Tien-Ping and Jia Jun, 2021). The work of Zhou et al. (2022) demonstrated how BART can be used for this purpose.

With respect to datasets for idiom paraphrasing, there are various mono-lingual English idiom datasets (Saxena and Paul, 2020; Zhou et al., 2021; Adewumi et al., 2021). Among these datasets, the PIE dataset (Zhou et al., 2021) stands out, as it contains both idiomatic expressions and their non-idiomatic counterparts.

When it comes to translation to another language rather than paraphrasing within a single language, Salton et al. (2014b) employed SMT to firstly substitute idioms with their simplified meanings before translation to the target language, after which the translated expressions were substituted with idioms in the target language.

Little research can be found when it comes to the direct translation of idiomatic expressions from a source to a target language. A major contributing factor to this could be the scarcity of paral-

lel corpora suitable for this task. A few papers on translation introduced their own datasets. An example is the work of Agrawal et al. (2018) where the authors introduced a parallel corpus of idiom-containing sentences in seven Indian languages and English, on which NMT and SMT models were trained. Fadaee et al. (2018) similarly created an English-German parallel corpus of sentences with idiomatic expressions, and evaluated the performance of NMT and SMT models based on it. Other corpora that support the development of idiom translation include a Russian-English (Aharodnik et al., 2018) and a Chinese-English dataset (Tang, 2022).

3 Methodology

The focus of this work is idiom translation. We thus consider idiom identification as outside of our scope, and make the assumption that input sentences contain idiomatic expressions. Furthermore, in some of our models (described below), the idiomatic expression itself is included as part of the input.

Below, we describe each of the two approaches we propose for idiom translation, one underpinned by idiom-to-idiom translation from the source to target language, and the other based on idiom paraphrasing within the source language followed by translation to the target language. In both cases, a T5 model was fine-tuned for a sequence-to-sequence learning task, where the input is provided in the form of a sequence of tokens and the model produces another sequence as its output. The task that the model needs to learn, is defined by prepending a prefix to the input sequence.

T5 models come in different sizes. In this work, we employed the `t5-small` implementation[1] which has around 60 million parameters and yet is feasible to train with limited computational resources. It comes pretrained for language modelling based on the Colossal Clean Crawled Corpus (Raffel et al., 2020) and fine-tuned for a number of downstream NLP tasks including translation.

3.1 Idiom-to-idiom Translation

In this approach, a model was developed to translate an idiom-containing sentence from the source language (English) to the target language (German).

[1] https://huggingface.co/t5-small

Dataset. The dataset used in training and evaluating our single translation model is the IdiomTranslationDS dataset by Fadaee et al. (2018). It consists of English-German sentence pairs where each of the sentences is an idiom-containing translation of the other. The idioms contained in each sentence pair are also provided. The dataset contains a total of 3498 sentence pairs sourced from the WMT training set (Bojar et al., 2017), distributed between a training and test set with 1998 and 1500 sentence pairs, respectively. Our analysis showed that certain idiomatic expressions appeared very frequently in this dataset. For instance, "to pass on", "to be in the know" and "in a nutshell" are contained in 513, 120 and 84 sentence pairs, respectively.

Data Cleaning. Manual inspection of the sentence pairs (carried out by two conversational German speakers) showed that some of the provided German sentences are not correct translations of their corresponding English sentences. To eliminate noise from the training and test sets, these pairs were manually removed, reducing the size of the training and test sets by 13.3% (265 pairs removed) and 15.2% (228 pairs removed), respectively. Furthermore, Unicode characters from other languages (Arabic and Mandarin) which appeared in some of the source and target sentences, were automatically removed.

As the original dataset did not provide predefined training and validation subsets, 15% of the pairs in the training set were randomly selected and held out to comprise a validation set.

Model Training. A T5 model was fine-tuned in different ways in order to develop different versions of our idiom-to-idiom translation model. This was carried out by varying the prefix prepended to the input sequence and/or specifying the pre-identified idiomatic expression within a given sentence.

The authors of T5 already provide a model that had been fine-tuned for a number of downstream tasks, including English-to-German translation (Raffel et al., 2020). As a starting point, the T5 model was further fine-tuned for the existing English-to-German translation task using our cleansed IdiomTranslationDS dataset. This required prepending the input sequences with the prefix *"translate English to German:"*. Additionally, we sought to define a new task for which to fine-tune T5, hence we trained another model whereby the input sequences were prepended with a custom

prefix, *"translate English to German with idiom:"*.

For both of the above fine-tuning tasks, we also investigated the effect of specifying the idiom contained within a given sequence. To this end, a suffix indicating the pre-identified idiom was appended to an input sequence. For example, the suffix *"idiom: to be in the picture"* was appended to the original input sequence *"She's not in the picture."* Four different translation models were obtained by fine-tuning the `t5-small` model for 50 epochs, for each of the following tasks: (1) Predefined translation: based on continuing to fine-tune T5 for the already existing English-to-German translation task, using the original input format; (2) Idiom-aware predefined translation: similar to task (1) but with the idiom appended at the end of the input sequence; (3) Custom translation: based on defining a new downstream task for T5, whereby we introduced the custom prefix *"translate English to German with idiom:"*; and (4) Idiom-aware custom translation: similar to task (3) but with the idiom appended at the end of the input sequence. Table 1 presents some examples that illustrate the different ways in which we fine-tuned the T5 model.

3.2 Idiom Paraphrasing and Translation

The second approach consists of a pipeline divided into two sub-tasks, each underpinned by a different model. The first sub-task is concerned with converting English idiom-containing sentences to their English paraphrases by training a paraphrasing model. This is followed by the second sub-task of translating the resulting paraphrases to German. It is worth noting that in the context of this approach, we define *paraphrase* as a simplification of the original idiom-containing sentence, allowing a reader to understand its meaning even if they are unfamiliar with the idiom. For example, a paraphrase of the sentence *"He feels he can paddle his own canoe after turning 18"* is *"He feels he can be self-reliant after turning 18."*

Dataset. To train the paraphrasing model, the PIE dataset (Zhou et al., 2021) was used. This dataset consists of English idiom-containing sentences as well as their corresponding paraphrases (also in English). Additionally, the dataset also specifies which tokens in a sentence corresponds to an idiom, as well as the meaning (sense) of that idiom. A total of 823 (non-unique) idioms are included in the dataset with a total of 5170 sentences, where each idiom has at least five sentence pairs per sense (as some idioms have multiple senses).

Data Cleaning. Although our analysis of the dataset showed that the data is mostly clean, several pre-processing operations were nevertheless applied. Some extraneous characters (e.g., ¾, ™) were removed. Also, variations in punctuation (e.g., different types of quotation marks) were normalised. Tokenisation seems to have been applied (by the dataset creators) on the data, e.g., "don't" appears as "do n't". However, this seems to be have been done inconsistently across the samples. Tokenised contractions were therefore merged again, considering that T5 does not require input sequences to be tokenised, as it comes with its own tokeniser.

The dataset was subdivided into training, validation and test sets following a 70-15-15% split.

Model training: Idiom paraphrasing. We explored a number of ways to fine-tune a T5 model for paraphrasing, while also exploiting the fact that a `t5-small` model fine-tuned specifically for general paraphrasing, is already available. This model, `t5-small-tapaco`[2], was fine-tuned on the TaPaCo dataset (Scherrer, 2020).

As our baseline, the original `t5-small` model was fine-tuned by introducing a new task specified by a custom prefix *"id_par:"* (short for "idiom paraphrasing"). This was necessary as none of the downstream tasks that T5 was originally fine-tuned for, were concerned with paraphrasing. To make the model aware of the idiom contained in a given sentence, we also appended the idiomatic expression itself, as supplied in the dataset.

Meanwhile, the `t5-small-tapaco` model which had already been fine-tuned for general paraphrasing, already recognises the predefined prefix *"paraphrase:"*. To fine-tune this specific model, we prepared the input sequences in our dataset by prepending the said prefix.

In summary, fine-tuning for the following tasks was performed (for 50 epochs), resulting in three types of paraphrasing models (exemplified in Table 2): (1) Custom paraphrasing with `t5-small`: based on fine-tuning `t5-small` whereby we introduced a custom prefix *"id_par:"* and appended the idiom at the end of the input sequence; (2) Predefined paraphrasing with `t5-small-tapaco`: based on conti-

[2] https://huggingface.co/hetpandya/t5-small-tapaco

Model Variant	Example Input Sequence
Predefined	*translate English to German: She's not in the picture*
Idiom-aware predefined	*translate English to German: She's not in the picture. idiom: to be in the picture*
Custom	*translate English to German with idiom: She's not in the picture.*
Idiom-aware custom	*translate English to German with idiom: She's not in the picture. idiom: to be in the picture*

Table 1: Examples showing how `t5-small` was fine-tuned for different tasks resulting in four translation model variants.

nuing to fine-tune `t5-small-tapaco` for paraphrasing, with the predefined prefix *"paraphrase:"* prepended to input sequences; and (3) Custom paraphrasing with `t5-small-tapaco`: based on fine-tuning `t5-small-tapaco` with the custom prefix *"id_par:"* and the idiom appended at the end of each input sequence.

Translation. The second sub-task is concerned with the translation of the English paraphrases (resulting from the first sub-task) to German. As the paraphrase model is presumed to have performed simplification of the idiomatic expressions contained in the input sentences, this sub-task can be cast as general translation from English to German. We leveraged the original `t5-small` model for this purpose, as it had already been fine-tuned for the English-to-German translation task.

4 Evaluation and Results

In order to evaluate our approaches, both automatic and human-based evaluation were conducted. Below, we first discuss the results of automatically evaluating each of the idiom-to-idiom translation and idiom paraphrasing models, followed by the results of human-based evaluation.

4.1 Automatic Evaluation

As part of our automatic evaluation, the following metrics were used: BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and COMET (Rei et al., 2020). COMET, in particular, has a variant known as Referenceless COMET, that we also used to estimate the quality of a generated translation even without a gold standard translation to compare with.

It is worth noting that an absolute score obtained by any of the above metrics is difficult to interpret on its own. Nevertheless, when viewed relative to each other, such scores are helpful in comparing the performance of different models and approaches (bearing in mind that for each of BLEU, METEOR and COMET, higher scores are desirable).

4.1.1 Idiom-to-idiom Translation

The BLEU, METEOR, COMET and Referenceless COMET scores obtained by our different idiom-to-idiom translation models on the cleansed IdiomTranslationDS test set are presented in Table 3. According to all metrics, the best results were obtained by the model that was based on continuing to fine-tune T5 for the predefined English-to-German translation task using the IdiomTranslationDS training set, without the idiomatic expression specified in the input sequence. To investigate whether the performance improvement obtained by this model (over the baseline model) is statistically significant, a paired t-test was performed for all scores. This resulted in p-values of 0.010, 0.056, 0.016 and 0.546 for BLEU, METEOR, COMET and Referenceless COMET, respectively. Considering a significance threshold of 0.05, we can say that the performance improvement based on BLEU and COMET is significant.

4.1.2 Idiom Paraphrasing and Translation

To evaluate the performance of our second approach, we firstly conducted a comparison of our different models for the idiom paraphrasing subtask. On the basis of that, the best-performing paraphrasing model was selected and integrated with our chosen English-to-German translation model to form a pipeline, whose performance was evaluated separately.

Shown in Table 4 are the results of evaluating our idiom paraphrasing models on our cleansed PIE validation set using the BLEU and METEOR metrics. Based on both scores, the best-performing paraphrasing model is the one that was based on fine-tuning `t5-small-tapaco` for our custom task using the cleansed PIE training set.

To realise the pipeline for our second approach, our Custom `t5-small-tapaco` model was integrated with the original `t5-small` model that was already fine-tuned for general English-to-German translation. This combination was evaluated on the cleansed PIE test set, the results of which are shown

Model Variant	Example Input Sequence
Custom `t5-small`	*id_par: The comedian had the audience in stitches. idiom: in stitches*
Predefined `t5-small-tapaco`	*paraphrase: The comedian had the audience in stitches.*
Custom `t5-small-tapaco`	*id_par: The comedian had the audience in stitches. idiom: in stitches*

Table 2: Examples showing how different paraphrasing model variants based on `t5-small` and `t5-small-tapaco` were fine-tuned based for different tasks.

Translation Model	BLEU	METEOR	COMET	Ref. COMET
Pretrained `t5-small` (Baseline)	0.145	0.493	0.241	0.100
Fine-tuned for Predefined task	**0.151**	**0.498**	**0.257**	**0.101**
Fine-tuned for Idiom-aware predefined task	0.146	0.495	0.255	0.097
Fine-tuned for Custom task	0.147	0.489	0.052	0.052
Fine-tuned for Idiom-aware custom task	0.142	0.492	0.242	0.095

Table 3: Evaluation results based on the cleansed IdiomTranslationDS test set. Referenceless COMET (Ref. COMET) scores were obtained by averaging over all test sentences. Each of the fine-tuned translation models is based on `t5-small`.

in Table 5. Compared to a baseline approach of translating an English idiom-containing sentence to German using the original `t5-small` model, our proposed pipeline-based approach obtained improved performance based on the Referenceless COMET metric. A paired t-test was performed and resulted in a p-value of 0.013, confirming that the improvement is statistically significant.

4.2 Human-based Evaluation

To complement the automatic evaluation carried out (described in Section 4.1), we sought the help of volunteer human participants in evaluating the outputs of our two approaches to idiomatic expression translation. To this end, a survey was built (using the Qualtrics platform[3]) to evaluate: (1) the extent to which each of our approaches generated fluent German sentences; and (2) how well each of our approaches generated German sentences that preserved the meaning of the original idiom-containing English sentences.

Survey design. The survey begins with a self-assessment section, which enabled us to ensure that responses were collected only from participants who are at least proficient/conversational in both English and German[4].

The core of the survey consists of two sections, each one intended to evaluate each of our two approaches. In each section, five questions were presented to a participant, where each question (described in more detail below) is intended to assess the quality of a generated German translation of an English idiom-containing sentence. Out of these five questions, two were fixed, i.e., shown to every participant, to allow us to calculate agreement between participants. The other three questions were based on random selection from a pool of 12 English idiom-containing sentences which were automatically translated by each of our approaches.

The first section was designed to evaluate the outputs of our best-performing idiom-to-idiom translation model. Each question presents an English idiom-containing sentence (drawn from the cleansed IdiomTranslationDS test set) and the German translation generated by the said model. A participant is asked to rate the translation in terms of fluency and meaning preservation on a scale of 1 to 5, with the values corresponding to: (1) "Very bad/Incomprehensible", (2) "Bad", (3) "Adequate", (4) "Good", and (5) "Very good/Flawless". An option for "I don't know" was also made available. An example question is shown in Appendix A.

The second section was designed similarly to the first section, except for the English idiom-containing sentences having been drawn from the cleansed PIE test set and their translations having been generated by our pipeline-based approach.

The survey, containing a total of 10 translation quality assessment questions spread across the two sections, was published for one week and obtained responses from a total of 53 participants.

[3] https://www.qualtrics.com

[4] We did not collect any personal information hence ethics approval of the survey was not required.

Paraphrasing Model	BLEU	METEOR
Custom `t5-small` (Baseline)	0.755	0.843
Predefined `t5-small-tapaco`	0.768	0.856
Custom `t5-small-tapaco`	**0.774**	**0.859**

Table 4: Results of evaluating our idiom paraphrasing models based on the cleansed PIE validation set.

	Paraphrasing		Translation
	BLEU	METEOR	Ref. COMET
Pretrained `t5-small` (Baseline)	NA	NA	0.0075
Custom `t5-small-tapaco`+Pretrained `t5-small`	0.768	0.852	**0.0147**

Table 5: Results of evaluating our combined idiom paraphrasing and translation approach, on the PIE test dataset. Referenceless COMET (Ref. COMET) scores were obtained by averaging over all test sentences.

Inter-rater agreement. In order to assess the reliability of the ratings collected through the survey, inter-rater agreement was calculated based on the fixed questions[5] that were presented to all participants. As a preliminary step, we removed any responses where the "I don't know" option was selected instead of a rating from 1 to 5, eliminating only two responses. We then calculated the value of Krippendorff's alpha (Hayes and Krippendorff, 2007) with the help of an implementation available from PyPi[6]. A value of 0.22 for alpha was obtained, which can be interpreted as fair agreement between our participants (Hughes, 2021). We do acknowledge that this implies that the rating task was not straightforward, and that a much higher agreement could have been obtained had we recruited only native German speakers (who also speak English), which we did not have access to at the time of this study.

Results. For each question in each section of the survey, the ratings given by participants were collected and analysed. The results for the idiom-to-idiom translation approach and the pipeline-based approach are visualised[7] in Figures 2 and 3, respectively. Each of the figures shows a box plot for every question, with the box ranging from the first to the third quartile and the whiskers extending to the minimum and maximum scores. It is worth noting that the set of 14 questions used in assessing the first approach (idiom-to-idiom translation) is different from the set of 14 questions used to assess the second approach (pipeline consisting of idiom paraphrasing and translation).

The median for every question is given by a vertical line, while the mean rating[8] is indicated by a star ($\star$). The dotted line represents the average score over all questions. The idiom-to-idiom translation model obtained an average fluency of 3.73, while that obtained by the pipeline-based approach is 3.65 (out of 5). In terms of meaning preservation, very similar average scores were obtained, i.e., 3.30 and 3.29 (out of 5) for the idiom-to-idiom translation and pipeline-based approaches, respectively.

5 Discussion

With respect to the first approach underpinned by idiom-to-idiom translation, our results showed that the best-performing model is the one that was based on continuing to fine-tune `t5-small` for the predefined English-to-German translation task. This shows that a T5 model that was fine-tuned for the predefined translation task, is better at translating English idiom-containing sentences to their idiom-containing German counterparts, compared to a model that was trained for a completely new idiom translation task. This is unsurprising considering that T5 was fine-tuned for general translation on the WMT 2014 English-German dataset with 4.5 million sentence pairs (Vaswani et al., 2017), while the cleansed IdiomTranslationDS dataset that we used to fine-tune T5 for the new, custom idiom translation task, includes only 1733 pairs.

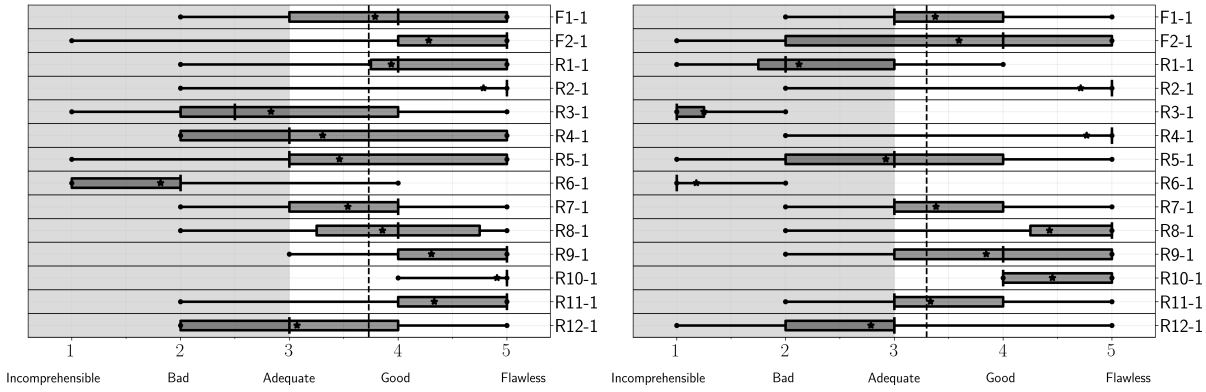
When it comes to the second approach which is based on a pipeline of idiom paraphrasing and translation models, our results demonstrate that fine-

[5]There were a total of four fixed questions given that two were included in each of the two sections.

[6]https://pypi.org/project/krippendorff/

[7]The box plots were produced based on code from https://github.com/mctenthij/CDS_paper

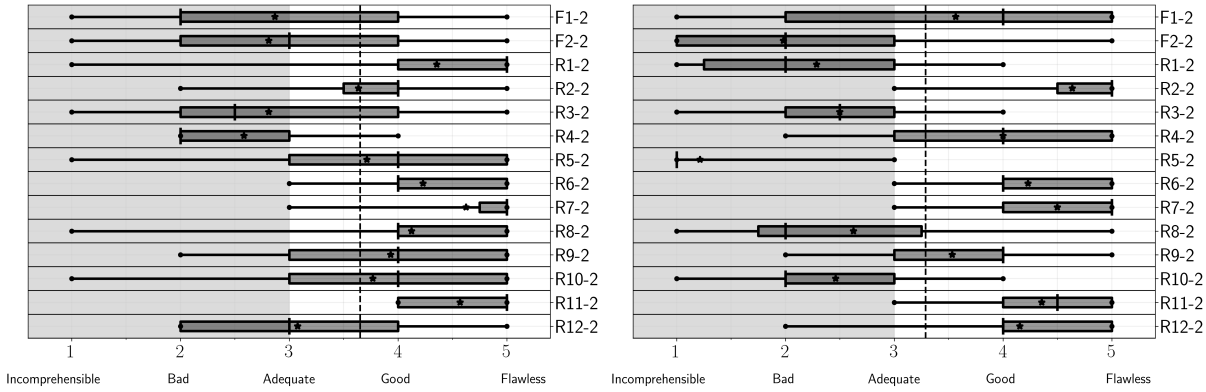
[8]The average number of ratings collected for the randomly selected questions is 13.25.



(a) Fluency scores

(b) Meaning preservation scores

Figure 2: Box plots of the scores provided by participants for each survey question assessing the quality of the outputs of our idiom-to-idiom translation model. Questions are denoted using the convention *F#-1* or *R#-1*, where *F* and *R* indicate a fixed and randomly selected question, respectively, and *1* means that the question was used to evaluate the first approach.



(a) Fluency scores

(b) Meaning preservation scores

Figure 3: Box plots of the scores provided by participants for each survey question assessing the quality of the outputs of the pipeline-based approach. Questions are denoted using the convention *F#-2* or *R#-2*, where *F* and *R* indicate a fixed and randomly selected question, respectively, and *2* means that the question was used to evaluate the second approach.

tuning `t5-small-tapaco` (a T5 model that had already been trained for general English paraphrasing) for our newly proposed custom paraphrasing task, leads to improved performance. Moreover, when this paraphrasing model is combined with the original T5 model for general English-to-German translation, better performance on the translation task is obtained, in comparison with using only the original T5 model.

Although the two approaches are not directly comparable with each other (i.e., the first approach is aimed at keeping the idiom in a German translation while the second one is aimed at generating a German translation of a non-idiomatic English paraphrase), our human-based evaluation shows that the first approach—the one based on idiom-to-idiom translation—seems to produce outputs which are marginally better that those of the second. This can be expected: as the second approach is based on a pipeline of paraphrasing and translation sub-tasks, any errors from the paraphrasing model would have been propagated to the translation model, affecting the quality of the final outputs. This is an issue that does not apply to the first approach since it performs direct translation.

6 Conclusions and Future Work

In this paper, we demonstrate how T5 models can be exploited in idiom translation: by fine-tuning them for idiom-to-idiom translation (first approach) and idiom paraphrasing (second approach). On the one hand, automatic evaluation showed that con-

tinuing to fine-tune the original T5 model for the predefined translation task on an idiom translation dataset, yielded optimal performance for idiom-to-idiom translation. On the other hand, fine-tuning a T5 model that had already been trained on a general paraphrasing task, for a custom idiom paraphrasing task, led to the best performance for idiom paraphrasing. Combining the said paraphrasing model with the original T5 model for general translation, resulted in improved results for idiom translation, compared with using just the latter. Human-based evaluation showed that both approaches produce translations of adequate quality.

To further advance research in idiom translation, we propose possible directions that can be pursued in the future. Firstly, a high-quality dataset with a much larger number of idiom-containing sentence pairs can be developed to facilitate better fine-tuning of T5 models for a custom idiom-to-idiom translation task. Moreover, it would be beneficial to create one dataset that can support the development of both idiom-to-idiom translation and idiom paraphrasing approaches. For instance, one can enrich the IdiomTranslationDS dataset by including paraphrases of idioms in English and German. Furthermore, our work has highlighted the fact that idiom translation datasets are scarce. When more such datasets become available, one can assess the extent to which our approaches can be applied to other language pairs.

To mitigate the current lack of large datasets for idiom translation, one could cast the idiom translation problem as a prompt-based learning task (Liu et al., 2021): a framework that makes it possible to apply pretrained language models to downstream tasks (such as translation) without the need for large amounts of data for fine-tuning.

References

Tosin P Adewumi, Saleha Javed, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. 2021. Potential idiomatic expression (PIE)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*.

Ruchit Agrawal, Vighnesh Chenthil Kumar, Vigneshwaran Muralidharan, and Dipti Misra Sharma. 2018. No more beating about the bush: A step towards idiom handling for Indian language NLP. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Katsiaryna Aharodnik, Anna Feldman, and Jing Peng. 2018. Designing a Russian idiom-annotated corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, Copenhagen, Denmark. Association for Computational Linguistics.

Rajesh Kumar Chakrawarti, Himani Mishra, and Pratosh Bansal. 2017. Review of machine translation techniques for idea of Hindi to English idiom translation. *International journal of computational intelligence research*, 13(5):1059–1071.

Verna Dankers, Christopher G Lucas, and Ivan Titov. 2022. Can Transformer be Too Compositional? Analysing Idiom Processing in Neural Machine Translation. *arXiv preprint arXiv:2205.15301*.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

GülŞen Eryiğit, Ali Şentaş, and Johanna Monti. 2022. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, page 1–33.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2018. Examining the Tip of the Iceberg: A Data Set for Idiom Translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

John Hughes. 2021. krippendorffsalpha: An R package for measuring agreement using Krippendorff's alpha coefficient. *arXiv preprint arXiv:2103.12170*.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Murathan Kurfali and Robert Östling. 2020. Disambiguation of potentially idiomatic expressions with contextual embeddings. In *Joint Workshop on Multiword Expressions and Electronic Lexicons, Barcelona, Spain (Online), December 13, 2020*, pages 85–94.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Changsheng Liu and Rebecca Hwa. 2016. Phrasal substitution of idiomatic expressions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 363–373, San Diego, California. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Qi Liu, Matt J Kusner, and Phil Blunsom. 2020. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014a. An empirical study of the impact of idioms on phrase based statistical machine translation of English to Brazilian-Portuguese. In *Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra)*, pages 36–41, Gothenburg, Sweden. Association for Computational Linguistics.

Giancarlo Salton, Robert Ross, and John Kelleher. 2014b. Evaluation of a substitution method for idiom transformation in statistical machine translation. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*, pages 38–42, Gothenburg, Sweden. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. EPIE dataset: a corpus for possible idiomatic expressions. In *23rd International Conference on Text, Speech, and Dialogue*, pages 87–94, Brno, Czech Republic. Springer.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

Caroline Sporleder and Linlin Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece. Association for Computational Linguistics.

Minghuan Tan and Jing Jiang. 2021. Does BERT understand idioms? A probing-based empirical study of BERT encodings of idioms. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1397–1407.

Kenan Tang. 2022. PETCI: A Parallel English Translation Dataset of Chinese Idioms. *arXiv*, abs/2202.09509.

Tan Tien-Ping and Dong Jia Jun. 2021. Translating idioms using paraphrasing, machine translation and rescoring. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(3):1942–1946.

University of Oxford. 2022. Oxford Learner's Dictionaries. Available online: `https://www.oxfordlearnersdictionaries.com/definition/english/idiom?q=idiom`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in Machine Translation. *Engineering*.

Xiao-mei Yu, Wen-zhi Feng, Hong Wang, Qian Chu, and Qi Chen. 2020. An attention mechanism and multi-granularity-based Bi-LSTM model for Chinese Q&A system. *Soft Computing*, 24(8):5831–5845.

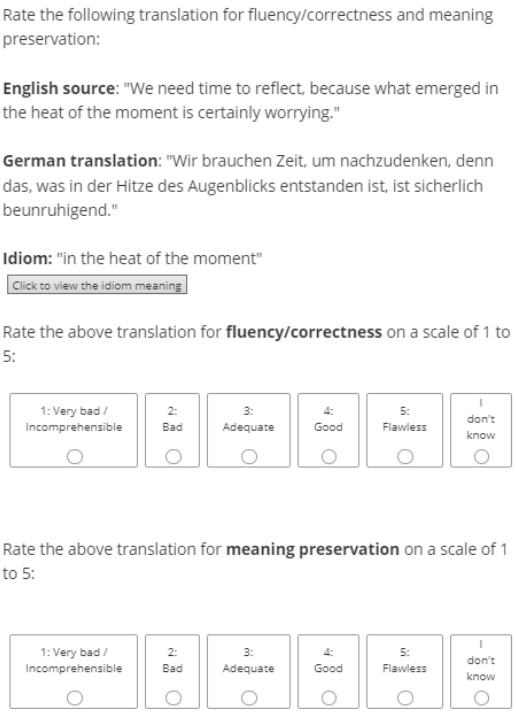
Jianing Zhou. 2021. Idiomatic sentence generation and paraphrasing. Master's thesis, University of Illinois.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

Jianing Zhou, Ziheng Zeng, Hongyu Gong, and Suma Bhat. 2022. Idiomatic Expression Paraphrasing without Strong Supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11774–11782.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. MICE: Mining Idioms with Contextual Embeddings. *Knowledge-Based Systems*, 235:107606.

A Appendix



An example of a survey question presented to participants as part of our human-based evaluation.