Evaluating Capabilities of MT Systems in Translating Idiomatic Expressions Using a Specialized Dataset

Steinþór Steingrímsson, Einar Freyr Sigurðsson
The Árni Magnússon Institute for Icelandic Studies
steinthor.steingrimsson@arnastofnun.is
einar.freyr.sigurdsson@arnastofnun.is

Abstract

Multiword expressions (MWEs) are generally problematic for machine-translation systems. In this paper, we (i) describe a set, available on CLARIN-IS, of appr. 1,000 idiomatic MWEs which have been translated into English; (ii) use the set as a template for a hidden evaluation set, to be used in a new leaderboard for Icelandic language technology, and (iii) evaluate – using both automatic and manual approaches – four MT systems' abilities to translate MWEs from Icelandic to English using both datasets. We find that traditional transformer-based MT systems evaluated commonly fail when translating idiomatic expressions, while LLMs do much better.

1 Introduction

Multiword expressions (MWEs) are a frequent phenomenon in natural language and speech. Proper handling of MWEs is important for various natural language processing (NLP) tasks, such as machine translation (MT), bilingual lexicon induction and information extraction. It is difficult to provide clear boundaries for what constitutes a MWE and what does not. The term can be used to describe fixed or semi-fixed phrases, compounds, idioms, phrasal verbs or collocations – in general, any sequence of words that acts as a single unit on some level (Calzolari et al., 2002).

In this paper, we introduce a set of approximately 1,000 Icelandic MWEs,[1] along with their translations into English as well as structured information about their usage. We consider all the MWEs in our dataset to be *idiomatic expressions*, i.e. idioms with an intended meaning that diverges from the literal meaning of the words constituting the expression, and therefore, they usually cannot be translated word for word.

Machine-translation systems generally do not handle MWEs well, and even though they are an important part of generating fluent translations, they can be a blind spot for traditional automatic evaluation approaches, such as BLEU (Papineni et al., 2002) or chrF++ (Popović, 2017). This applies especially in cases where there is more than one "right" answer, as the traditional lexical metrics cannot identify what goes wrong in a translation. The Icelandic MWE dataset was compiled for use with MT, and can be used either to augment training sets with sentence pairs containing common idiomatic expressions, or for evaluating the capabilities of MT systems to translate such expressions. We show how the dataset can be used to evaluate the capabilities of four MT systems to translate MWEs, by evaluating the systems in three different ways: using traditional automatic approaches, using automatic evaluation of MWE translations, and by manually evaluating the output.

For an accurate evaluation of an MT system capabilities, the evaluation data should not be used for training the system. Large language models (LLMs), which demand enormous amounts of data, are primarily trained on web-crawled data. When a dataset is made available online for a prolonged period of time, it becomes increasingly likely that it has been gobbled by spiders crawling the web for LLM training data. Therefore, it can be said that openly available evaluation sets start to rot, meaning that they become increasingly untrustworthy, as soon as they are put online. To still be able to gauge the capabilities of MT systems in translating MWEs from Icelandic into English, we tackle this problem by using the publicly available dataset from the CLARIN-IS repository as a template for another dataset, Hidden Idiomatic Expressions for Machine Translation Evaluation (HIDEMATE). In order to hinder data

[1] http://hdl.handle.net/20.500.12537/275 (Halldórsson et al., 2022).

leakage, HIDEMATE will not be made publicly available and will only be used as a part of a machine translation (MT) evaluation suite to be deployed on a leaderboard for Icelandic language technology, set to open in the fall of 2025. The dataset will be controlled by the Árni Magnússon Institute for Icelandic Studies (AMI), stored on internal servers and evaluations using the datasets will only be carried out by AMI staff.

In the paper we will describe how we evaluate MT systems using these datasets as well as carrying out initial evaluations. Section 2 discusses related work. In Section 3 we briefly describe the MWE datasets whereas Section 4 discusses MT evaluation. We demonstrate our results in Section 5, discuss future work in Section 6 and conclude in Section 7. In Section 8, we discuss potential limitations of this work.

2 Related Work

In recent years, idiomatic expressions (and MWEs in general) have been the focus of much work in natural language processing (NLP), not least in MT research. Stap et al. (2024) introduce a dataset containing one thousand idioms in context with translations for three translation directions, English→German, German→English and Russian→English, and use it to evaluate LLMs fine-tuned using their fine-tuning approach. Tang (2022) present a parallel English dataset of Chinese idioms, and Ármannsson et al. (2024) introduce a dataset for evaluating English→Icelandic idiomatic expressions and an approach for how to go about the evaluation.

In Ármannsson et al.'s (2024) submission to the WMT24 test suite subtask, they focus on evaluating the capabilities of participating models in translating idiomatic expressions from English to Icelandic. This is the first published attempt to systematically compare translational capabilities of automatic systems working with Icelandic in terms of MWEs. They use two lists of words for each idiom, one with literal translations, which are a negative match when translating idiomatic expressions, and one with a positive match, which contains the words of the most likely idiomatic expression. In our work we use this latter type of list, but not the former. The idiom *taka <einhvern> til bæna*, lit. 'take <someone> to prayers', can be translated at least as 'read <someone> the riot act' and 'take <someone> to task'. For a successful translation, according to our automatic evaluation, we require the translation to either include all of the words *read*, *riot*, *act* or all of the words *take*, *to*, *task*.

3 Collecting the Multiword Expressions

The set of multiword expressions, distributed on the Icelandic CLARIN repository,[2] contains approximately 1,000 Icelandic idioms processed from the ISLEX dictionary (Úlfarsdóttir, 2014). They are listed with their English idiomatic equivalent and literal meaning in both languages, as well as example sentences and keywords. The idioms are, in most cases, syntactically mobile, which is why case information is included.

The idioms were processed from a list of 4,000 MWEs in the ISLEX database. The idioms are ordered alphabetically according to the first keyword of each idiom and each line contains the following categories: 1) Icelandic idiom; 2) English equivalent; 3) Meaning of the Icelandic idiom; 4) Meaning of the English idiom; 5) An example sentence with the Icelandic idiom; 6) An example sentence with the English idiom; 7) An example sentence with the meaning of the Icelandic idiom; 8) An example sentence with the meaning of the English idiom; 9) Keywords in the Icelandic idiom, lemmatized (in some cases in the plural). Table 1 exhibits an example from the dataset where all nine columns are shown. The first four categories contain type information, cf. for example, the idiom *rétta <einhverjum> hjálparhönd* 'lend <someone> a helping hand', which is listed as follows: <NP1-nom> rétta <NP2-dat> hjálparhönd; <NP1> lend <NP2> a helping hand; <NP1-nom> hjálpa <NP2-dat>; <NP1> help <NP2>.

Where more than one English equivalent, translation or sense are possible, alternatives are separated with a pipe symbol. Keep in mind that there is not always a 1-1 relation between the example sentences. For example, the Icelandic idiom *Það er fokið í flest skjól* is translated as 'We're at the end of our tether', where the Icelandic expletive *það* 'it, there' makes way for a personal pronoun in English. The use of other symbols in the file is as follows: alternatives within the same segment are separated with a slash (/),

[2] https://repository.clarin.is/repository/xmlui/handle/20.500.12537/275

Source idiom	**Target idiom**
<NP1-nom> rétta <NP2-dat> hjálparhönd	<NP1> lend <NP2> a helping hand
Source sense	**Target sense**
<NP1-nom> hjálpa <NP2-dat>	<NP1> help <NP2>
Source idiomatic example	**Target idiomatic example**
Sigurður rétti Guðmundi hjálparhönd	Sigurður lent Guðmundur a helping hand
Source sense example	**Target sense example**
Sigurður hjálpaði Guðmundi	Sigurður helped Guðmundur
Keywords	
hjálparhönd	

Table 1: An example of an idiom in our dataset.

as in, e.g., the idiom *vera klár/tilbúinn í slaginn* ('be ready to rumble'), and optional parts of idioms are in parentheses, e.g. the idiom *bretta upp ermar(nar)* ('roll up one's sleeves') or *vera sjálfs sín(s) herra* ('be one's own boss').

There are a few examples of duplicate lines in the file with respect to the source idiom, but only in cases where the respective meaning can be considered twofold, as for example in the idiom *ganga ekki heill til skógar*, which can either refer to physical or mental health, i.e. 'be under the weather' (physical) or 'not be playing with a full deck' (mental).

Users of the dataset will note that the Icelandic male names *Sigurður* and *Guðmundur* are used as actors in the example sentences. This is for the sole reason that they have different inflectional forms for each case (nom. *Sigurður/Guðmundur*, acc. *Sigurð/Guðmund*, dat. *Sigurði/Guðmundi*, gen. *Sigurðar/Guðmundar*).

For the MT evaluation, we process the data in a slightly different way than in the distribution file. We number each segment, and while we only use the example phrases and their translations, where there are alternatives within the segments we generate all possible pairs. The generated pairs then get the segment number and the evaluation results are weighted so that all segments in the dataset have the same weight in the final score. Furthermore, we add a list of words that should be included in the MT translation of the idiom, and that list is used for the automatic evaluation of idiom translation. The processed data, along with all scripts, are made available on GitHub.[3]

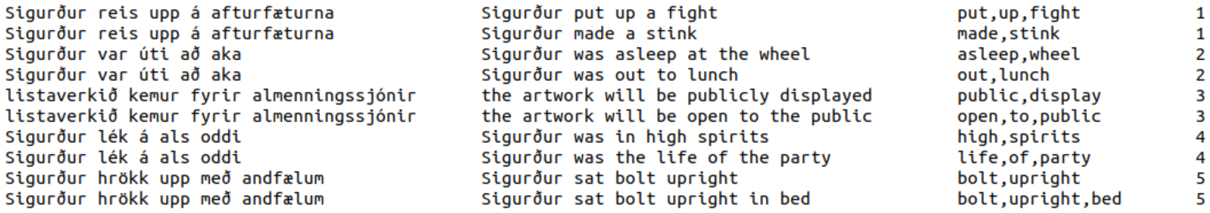
HIDEMATE (Hidden IDiomatic Expressions for MAchine Translation Evaluation) is processed in the same way and only contains data fields necessary for evaluation. Figure 1 shows examples of how the dataset from the CLARIN-IS repository is prepared for evaluation. The first column contains the source sentence in Icelandic. The second column contains a reference translation in English. Note that for each source sentence there can be one or more reference translations. The keywords that should be included in an MT translation of the idiom are in column 3. Column 4 has an id for the source sentence, as sometimes the same source sentence has multiple translations.

4 Evaluating Machine-Translation Systems

When choosing which MT system to use for a given task, the ability to translate MWEs can be a deciding factor. When we evaluate MT systems' ability to translate idioms from one language to another, we may want to punish incorrect literal translations. An example would be if an MT system would translate *kick the bucket*, meaning 'die', from English to, say, Icelandic and we would find both the word *sparka* 'kick' and *fata* 'bucket' as *sparka í fötuna* can only mean that someone strikes a bucket with their feet.

It is therefore important to be able to test these capabilities. To this end, we run three evaluation

[3] https://github.com/stofnun-arna-magnussonar/IdiomaticExpressions



Sigurður reis upp á afturfæturna	Sigurður put up a fight	put,up,fight	1
Sigurður reis upp á afturfæturna	Sigurður made a stink	made,stink	1
Sigurður var úti að aka	Sigurður was asleep at the wheel	asleep,wheel	2
Sigurður var úti að aka	Sigurður was out to lunch	out,lunch	2
listaverkið kemur fyrir almenningssjónir	the artwork will be publicly displayed	public,display	3
listaverkið kemur fyrir almenningssjónir	the artwork will be open to the public	open,to,public	3
Sigurður lék á als oddi	Sigurður was in high spirits	high,spirits	4
Sigurður lék á als oddi	Sigurður was the life of the party	life,of,party	4
Sigurður hrökk upp með andfælum	Sigurður sat bolt upright	bolt,upright	5
Sigurður hrökk upp með andfælum	Sigurður sat bolt upright in bed	bolt,upright,bed	5

Figure 1: Examples of sentences prepared for evaluation.

experiments. First, we simply evaluate the MT output using traditional automatic approaches. We apply the common evaluation metrics BLEU, chrF++ and COMET. Second, we devise a simple automatic approach that classifies translations in two groups: translations likely to have correctly handled the MWE and translations that failed to do so. Third, we manually evaluate all translations to be able to confirm or reject the adequacy of the automatic approach.

4.1 MT Systems

We compare four systems capable of Icelandic to English machine translation: two LLMs and two dedicated translation systems. All these systems are publicly available and as of early 2025 are all commonly used by the Icelandic-speaking population. These systems are OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet, Google Translate and the translation system on M.is.

4.1.1 GPT-4o

ChatGPT caught the world by storm in late 2022. The chatbot was built on a large language model, GPT-3, taking advantage of reinforcement learning with human feedback (RLHF) to create impressive conversational abilities. Since then multiple improvements have been made as well as different versions of the underlying language model. Among the many uses of the system is automatic translation, as the underlying LLMs are trained on data in multiple languages and have cross-lingual capabilities. In our experiment we use GPT-4o, which the OpenAI website states is their "versatile, high-intelligence flagship model". [4] We employ the default GPT-4o version at the time of our experiment, `gpt-4o-2024-08-06`.

We accessed the model through API to carry out a three-shot translation. The examples given to the system were the ones used to collect translations from LLMs for the WMT 24 general translation task (Kocmi, Avramidis, Bawden, Bojar, Dvorkovich, Federmann, Fishel, Freitag, Gowda, Grundkiewicz, Haddow, Karpinska, Koehn, Marie, Murray, et al., 2024).[5] We set the temperature to a rather low number, 0.2, for more deterministic output. The request template is shown in Figure 2.

4.1.2 Claude 3.5 Sonnet

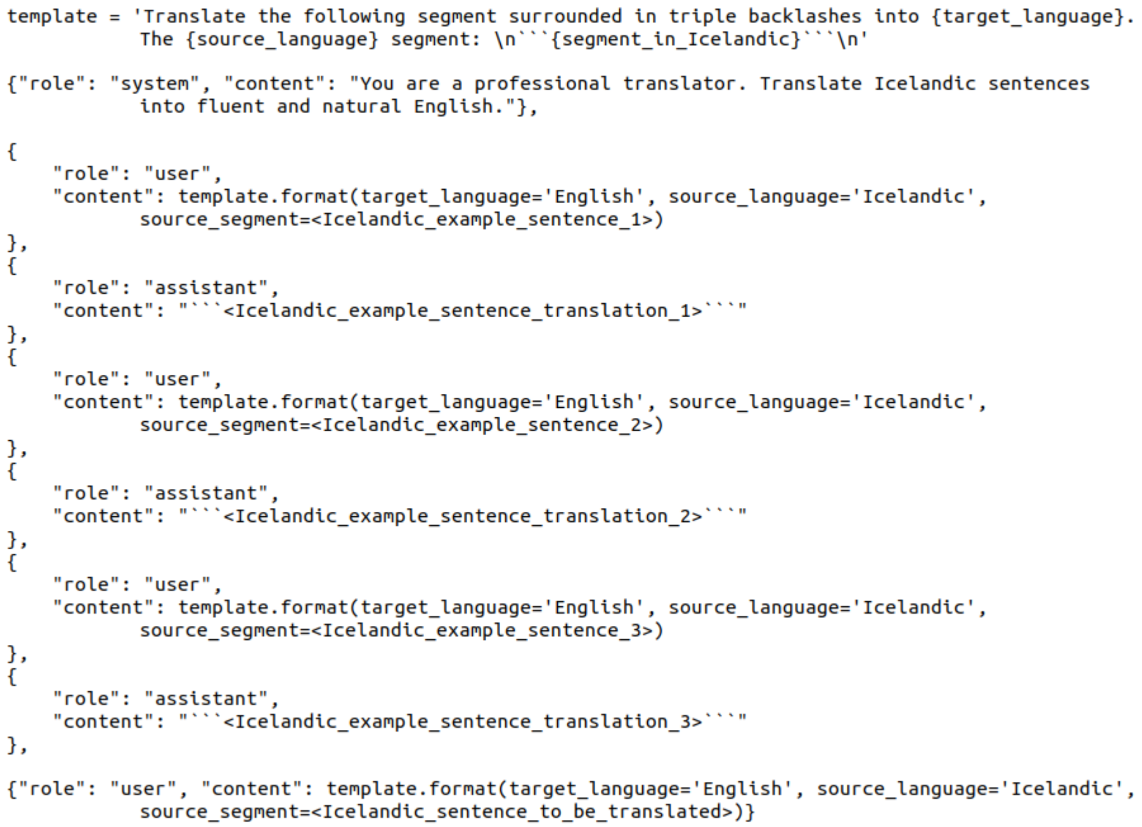
As well as being widely used, Anthropic's Claude 3.5 was the highest scoring openly available system in the WMT 24 General Translation Task (Kocmi, Avramidis, Bawden, Bojar, Dvorkovich, Federmann, Fishel, Freitag, Gowda, Grundkiewicz, Haddow, Karpinska, Koehn, Marie, Monz, et al., 2024) for the English→Icelandic translation direction. It is thus an obvious choice for our MWE translation evaluation for the opposite direction, Icelandic→English.

We employ Claude 3.5 Sonnet, which is the "most intelligent model" according to Anthropic's website.[6] For our experiments we select the default Claude 3.5 Sonnet version at the time of our experiment, `claude-3-5-sonnet-20241022`. As with the OpenAI model we set the temperature to 0.2, and format the request template in the same way as before, shown in Figure 2.

[4] https://platform.openai.com/docs/models#gpt-4o, accessed on February 6th, 2025.

[5] Made available here: https://github.com/wmt-conference/wmt-collect-translations

[6] https://docs.anthropic.com/en/docs/about-claude/models, accessed on February 6th, 2025.



````
template = 'Translate the following segment surrounded in triple backlashes into {target_language}.
           The {source_language} segment: \n```{segment_in_Icelandic}```\n'

{"role": "system", "content": "You are a professional translator. Translate Icelandic sentences
           into fluent and natural English."},

{
    "role": "user",
    "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_example_sentence_1>)
},
{
    "role": "assistant",
    "content": "```<Icelandic_example_sentence_translation_1>```"
},
{
    "role": "user",
    "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_example_sentence_2>)
},
{
    "role": "assistant",
    "content": "```<Icelandic_example_sentence_translation_2>```"
},
{
    "role": "user",
    "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_example_sentence_3>)
},
{
    "role": "assistant",
    "content": "```<Icelandic_example_sentence_translation_3>```"
},

{"role": "user", "content": template.format(target_language='English', source_language='Icelandic',
           source_segment=<Icelandic_sentence_to_be_translated>)}
````

Figure 2: The template for the requests to gpt-4o and Claude 3.5 Sonnet APIs.

4.1.3 Google Translate

For many people, Google Translate has been the go-to MT system for Icelandic since the language was added in 2009. Google Translate is fast and accessible and widely used. It has commonly been used by the Icelandic MT research community to compare their models to, ever since the first paper on Icelandic MT using statistical and neural methods was published by Jónsson et al. (2020). For our experiment we access Google Translate through an API.[7]

4.1.4 m.is

m.is[8] is a dictionary and language technology portal, which opened in mid-2024 and is hosted by AMI. It is aimed at students of Icelandic, native speakers of Icelandic, as well as second language learners. The website gives access to a contemporary dictionary of Icelandic, a database of inflections, Icelandic→English and Icelandic→Polish bilingual dictionaries as well as an MT system translating both ways between Icelandic and English.

The MT system on m.is is based on the submission to the WMT24 general news translation task by Jasonarson et al. (2024). While the WMT task was only to translate English→Icelandic, a model was also trained in the other direction, Icelandic→English, to generate backtranslations for training the final system.

The submission was based on four transformer (Vaswani et al., 2017) models of various sizes, with each of them generating multiple candidates. The final translation is selected using COMET (Rei et al., 2020). For the backtranslations from Icelandic to English only one model was trained, a Transformer$_{\text{BIG}}$ model of approximately 200M parameters. This model was then deployed to carry out

[7] Google Translate was used to translate the sentences on February 6, 2025.

[8] https://m.is

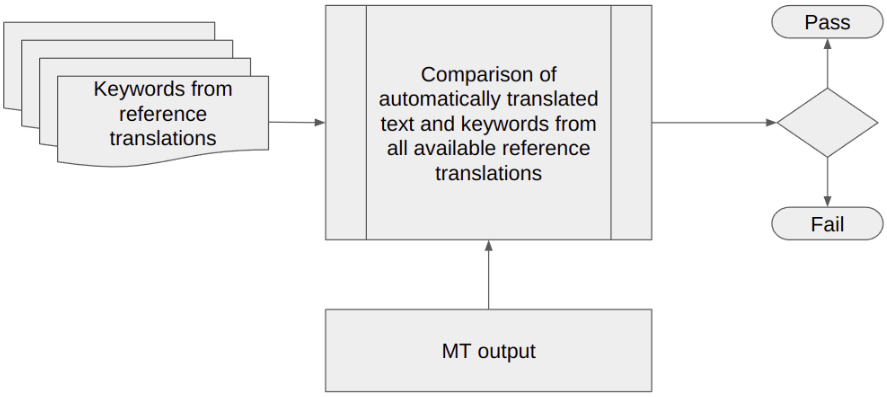

Keywords from reference translations
Comparison of automatically translated text and keywords from all available reference translations
Pass
Fail
MT output

Figure 3: The AutoIE process.

Icelandic→English translations on m.is, by generating five candidates, and having COMET selecting the best one for output.[9]

4.2 Automatic Evaluation Approaches

In order to make a general comparison of the MT systems used, we calculated BLEU, chrF++ and COMET scores for translations of all sentences in the two datasets. The scores for BLEU[10] and chrF++[11] were calculated using Sacrebleu (Post, 2018). Sacrebleu signatures are given in footnotes and results reported in Table 2 for the dataset from the CLARIN-IS repository and in Table 4 for HIDEMATE.

4.2.1 Automatic Idiom Evaluation

For our experiments, we devised a simple automatic approach, designated AutoIE (Automatic Idiom Evaluation), to gauge how well the MT systems managed to process the idiomatic expressions. Each machine-translated output is assigned a pass or a fail. The translation gets a pass if it contains all content words of the translation in the dataset. For example, for the sentence *Sigurður fékk sér kríu*, translated in the dataset as ‘Sigurður took a nap’, the MT translation has to contain the words ‘took’ and ‘nap’ to receive a pass. If it does not contain both words, it is assigned a fail. Figure 3 shows how the AutoIE algorithm is passed a translation candidate from an MT system and provided with one or more reference translations and keywords for each one. If all the keywords for any of the reference translations is found in the MT translation candidate, the system assigns a pass. Otherwise it assigns a fail.

Note that each source sentence can have multiple different acceptable translations, but they need to be defined in the reference set. The reference file for the dataset from the CLARIN-IS repository has been made available in the project’s GitHub-repository.[12]

4.2.2 BLEU

The BLEU score, introduced in 2002 (Papineni et al., 2002), was the de facto standard for MT Evaluation for twenty years and only in the early 2020s have other metrics come to topple its dominance, with 2024 being the first year that BLEU was not used as one of the evaluation metrics for the WMT general translation task. BLEU compares an automatically translated text to one or more human-written reference translations. It does so by checking how many n-grams in the candidate translation appear in the

[9] The m.is translation engine was used to translate the sentences on February 6, 2025.

[10] BLEU|nrefs:3|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1

[11] chrF2|nrefs:3|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

[12] https://github.com/stofnun-arna-magnussonar/IdiomaticExpressions

reference translation, with the default being all n-grams up to 4-grams. It also applies a brevity penalty for translations that are much shorter than the reference.

While BLEU is easy to use and can give a good idea of the likeliness of a candidate translation and the reference, over the years it has been criticized for only measuring surface features while not considering meaning or context. It also does not evaluate syntax, grammar or sentence fluency and cannot evaluate long-range relationships between words. It depends on reference translations, and a single reference may not cover all possible valid translations as BLEU does not account for synonyms or paraphrases.

These drawbacks often make it hard to use BLEU to evaluate translations rich in idiomatic expressions as they are likely to have multiple valid translations, which can be idiomatic or literal in some cases, but not in others. In our experiment we will calculate BLEU scores for the translations and compare it to the AutoIE score as well as manual evaluation.

4.2.3 chrF++

chrF++ is another surface-level evaluation metric for MT. It measures precision and recall of character n-grams, which makes it more suitable when translating into morphologically rich languages, such as Icelandic. chrF++ also considers word n-grams, which improves its correlation with human judgment, but it is not as strict as BLEU in penalizing word reordering. Like BLEU it has multiple drawbacks related to it not considering meaning or context or properties appertaining to that.

4.2.4 COMET

Unlike the traditional n-gram based metrics, COMET uses neural models to compare an automatically generated translation to the source sentence and a reference translation. These inputs are processed by multilingual transformer models to generate semantic representation. These models have been trained on human judgments of MT quality, such as the publicly available human evaluations from the WMT evaluation campaigns pertaining to the general translation tasks which take place each year.

COMET has come to be quite popular in recent years, especially when evaluating language pairs for which there are substantial resources of training data. Multiple different COMET models have been made available. For evaluating our translation experiments we use wmt22-comet-da.[13]

4.3 Manual Evaluation

To assess whether our automatic approach is useful a sample of translations from each set, 220 sentences from the dataset from the CLARIN-IS repository and 220 sentences from HIDEMATE, for each of the four MT systems, were evaluated by a linguist whose task was only to look at the MWE and assess whether it was translated correctly. The evaluator would select one of three options: *Correct translation*, *Incorrect translation* and *Unusual translation but can be understood*. It can be difficult to determine whether a translation is correct and this last category, *Unusual translation but can be understood*, was sometimes used if the evaluator was uncertain, for example, if he found a translation to be fairly good but he did not feel it captured the meaning perfectly. Also in this category were translations where the meaning was captured but the wording was not entirely what one would expect. The linguist was only to look at the MWE and disregard all other possible errors in the translation. The results for the manual evaluation for the CLARIN-available dataset are given in Table 3 and the manual evaluation for HIDEMATE in Table 5.

Upon inspecting the outputs, we find that idioms that can be translated word by word from Icelandic into English, such as *Sigurður var úlfur í sauðargæru* (‘Sigurður was a wolf in sheep’s clothing’) and *Sigurður bjargaði andlitinu* (‘Sigurður saved face’; lit. ‘Sigurður saved **the** face’), are most likely to be translated correctly. Idioms that require translating into an idiom that has the same meaning but uses a different metaphor are less likely to be translated correctly. Example of that could be *Sigurður er eldri en tvævetur*, literally ‘Sigurður is older than two winters old’, which would normally be translated into ‘Sigurður was not born yesterday’, or an idiom containing words where the most common sense is not the one carried in the idiom, such as *Sigurður rak lestina* (‘Sigurður trailed behind’) which contains the word *lest*, perhaps most commonly meaning a locomotive train and translated as ‘train’.

[13] https://huggingface.co/Unbabel/wmt22-comet-da

MT System	BLEU	chrF++	COMET	AutoIE (%)
gpt4-o1	29.0	57.7	0.6559	25.3
Claude 3.5 Sonnet	29.9	53.0	0.6572	26.3
Google Translate	19.4	51.6	0.6166	15.8
m.is	23.2	52.0	0.6434	14.8

Table 2: Automatic evaluation of the four MT systems on the CLARIN-available dataset.

MT System	Correct (%)	Understandable (%)	Incorrect (%)
gpt4-o1	68.5	9.6	21.9
Claude 3.5 Sonnet	69.0	8.7	22.3
Google Translate	35.6	7.8	56.6
m.is	38.4	9.6	52.0

Table 3: Manual evaluation of the four MT systems on the CLARIN-available dataset.

MT System	BLEU	chrF++	COMET	AutoIE (%)
gpt4-o1	34.9	61.3	0.6867	35.5
Claude 3.5 Sonnet	36.8	57.9	0.6861	34.1
Google Translate	25.1	54.8	0.6405	21.8
m.is	26.3	52.9	0.6494	20.0

Table 4: Automatic evaluation of the four MT systems when translating HIDEMATE.

MT System	Correct (%)	Understandable (%)	Incorrect (%)
gpt4-o1	81.1	6.0	12.9
Claude 3.5 Sonnet	80.7	7.8	11.5
Google Translate	44.0	6.5	49.5
m.is	44.2	10.2	45.6

Table 5: Manual evaluation of the four MT systems on the HIDEMATE dataset.

5 Results

The LLMs scored highest on all the metrics, usually by a large margin. In the automatic evaluation, they translated over 25% of the idiomatic expressions in the CLARIN-available dataset correctly (see AutoIE in Table 2) and more than a third in the HIDEMATE dataset (see AutoIE in Table 4). We can see that there is a slight variation in the order of the systems by which metric is used. Claude always obtains the highest BLEU scores, while GPT4-o1 obtains the highest chrF++ scores. Similar variations can be seen when we compare Google Translate and m.is, with m.is scoring higher on BLEU and COMET, but Google Translate obtaining higher chrF++ scores on the HIDEMATE dataset and a very close chrF++ score on the CLARIN-available dataset. Google Translate also scores higher on the AutoIE metric.

This shows that the choice of evaluation metric is important when it comes to selecting an MT system for a given task. While all the metrics give the general picture, there is a considerable variation in how close the scores are between two system. We also see that when the traditional metrics show a difference between two systems, like for m.is and Google Translate in our experiment, this does not necessarily indicate that the "better" system is more adept in translating something like idiomatic expressions.

The results in the manual evaluation demonstrate that the LLMs get a lot higher score than Google Translate and m.is, which are both based on encoder-decoder models. gpt4-o1 is not far behind Claude3.5 Sonnet in the CLARIN-available dataset but scores slightly higher in the HIDEMATE dataset. In both cases though, the difference is negligible. While Google Translate and m.is translate considerably fewer idioms correctly than the LLMs, there is little difference between the two. m.is obtains a slightly higher number of correct translations for both datasets, but the difference between the two is more pronounced in the number of incorrect translations, where Google Translate does rather worse than m.is.

When we compare the manual evaluation to the results of AutoIE, we find that the automatic metric only accepts approximately half of what the human evaluator accepts. However, even though it does not give more accuracy than that, its failings seem to apply to all datasets and all MT systems, and gives very similar results, although Google Translate scores slightly higher than m.is on AutoIE, while m.is scores slightly higher on the manual evaluation. Overall the results indicate that the AutoIE metric gives useful information on the capabilities of translation systems when translating idiomatic expressions.

Our experiments also show that LLMs are considerably better than smaller transformer models in translating idiomatic expressions. The LLMs do produce less literal outputs, compared to the transformer models and this seems to be particularly true when dealing with idiomatic expressions. Our results are in line with previous research, for example Ármannsson et al. (2024), which found that while the best transformer-based models could be equally good or even better than LLMs in cases where a literal translation was called for, they were no match when it came to idiomatic expressions.

Finally, even though we do not intend to publish the HIDEMATE evaluation set or making it accessible in any way, it is important to be able to describe its content and give examples of the kind of data that is included in it. We modeled HIDEMATE on a dataset of idiomatic expressions that has been made available on CLARIN and ideally the two datasets should give very similar scores when evaluating the same system. While both datasets give results close enough to give the same general idea, there is quite some difference between the scores.

6 Future Work

In some cases, each Icelandic example is given only one translation in our datasets, although more translations may be valid. Adding additional valid translations for each example would be useful in order for automatic evaluations to align more closely to manual evaluation scores when the datasets are used with the AutoIE approach to evaluate the capabilities of MT systems to translate idiomatic expressions. By comparing the translations deemed correct in the human evaluation to the translations given in the datasets, we can add more valid translations. We intend to do so for future versions of the datasets in order to make them even more viable for automatic evaluation.

We have not divided our datasets into different types of MWEs, depending on how transparent or opaque they are. However, that might be helpful for determining what kind of idioms the MT systems are best at translating. For some opaque idioms, all the MT systems fail. An example is *Sigurður dró ýsur* which means that Sigurður nodded off or dozed but the MT systems translate it literally, as ‘Sigurður pulled/caught haddocks’. By categorizing the idioms with respect to transparency, we can make sure that the different datasets we are evaluating are as similar to each other as possible.

Our results show that there is great enough difference in two datasets to result in quite different scores. We want to investigate further why that is, and in a later version try to take care to remedy this difference.

Finally, we intend to use the openly available dataset introduced here as a supplemental data for MT training, and investigate if that will increase the capabilities of an MT system to translate idioms, as measured by our hidden dataset, HIDEMATE.

7 Conclusions

The evaluation results, and the result analysis, indicate that the traditional MT systems commonly fail when translating idiomatic expressions, while LLMs usually do not. Specialized evaluation sets, such as the one introduced in this paper, can be used to gauge the capabilities of different systems. The simple automatic approach introduced here provides results in line with a thorough manual evaluation, indicating that it may be sufficient to help in the selection of the best system in this regard, when needed.

8 Limitations

There are various potential limitations to our work and here we mention a few of them.

While we want to prevent data leakage by building a hidden evaluation set, HIDEMATE, we process the idioms from ISLEX which is available online. It could be the case that the ISLEX data is among the web-crawled texts in the LLMs’ training data. While that data is not formatted to be made available in sentence pairs, as an evaluation set is, the proximity of idioms to their explanations and translations in the online dictionary could help LLMs learn these exact idioms when trained on the web-scraped dictionary data.

It is not always clear cut what constitutes an MWE which is not an idiom and what is clearly an idiom. There may be a gray area there. When compiling our datasets we select the MWEs which we consider idiomatic, in some cases others might disagree.

In our manual evaluation, all four translations for a given sentence appear in the same order. This could lead to a bias if the worst model or models do something unusual while the best models do not, as the evaluator may lean towards what he can expect beforehand.

AutoIE is a word inclusion test, testing for a non-literal translation of idioms. It is potentially prone to recognising presence of the proper translation while, in fact, the words in the target language may not form a proper expression. While it does not seem to our evaluator to be a common phenomenon, estimating how common it is could give us more insight into the feasibility of our automatic approach.

Finally, while the evaluator is a native speaker of Icelandic and speaks English fluently, he is not a native speaker of English. That is a clear limitation when judging whether translations into English are correct or not. However, as he was the only evaluator, there should be internal consistency in his judgments.

Acknowledgments

We thank three anonymous reviewers for helpful comments which improved the paper. The making of the CLARIN-available dataset (Idiomatic Expressions (Icelandic and English) 22.09) was funded by the Ministry of Education, Science and Culture in Iceland as part of the Language Technology Programme for Icelandic 2019–2023. We thank our collaborators who worked on the dataset: Björn Halldórsson, Árni Davíð Magnússon and Finnur Ágúst Ingimundarson.

References

Ármannsson, B., Hafsteinsson, H., Jasonarson, A., & Steingrímsson, S. (2024). Killing Two Flies with One Stone: An Attempt to Break LLMs Using English-Icelandic Idioms and Proper Names. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 451–458). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.31

Calzolari, N., Fillmore, C. J., Grishman, R., Ide, N., Lenci, A., MacLeod, C., & Zampolli, A. (2002). Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the Third International Conference on Language Resources and Evaluation* (pp. 1934–1940). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2002/pdf/259.pdf

Halldórsson, B., Magnússon, Á. D., Ingimundarson, F. Á., Sigurðsson, E. F., Steingrímsson, S., Jónsdóttir, H., & Úlfarsdóttir, Þ. (2022). Idiomatic Expressions (Icelandic and English) 22.09 [CLARIN-IS]. http://hdl.handle.net/20.500.12537/275

Jasonarson, A., Hafsteinsson, H., Ármannsson, B., & Steingrímsson, S. (2024). Cogs in a Machine, Doing What They're Meant to Do – the AMI Submission to the WMT24 General Translation Task. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 253–262). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.18

Jónsson, H. P., Símonarson, H. B., Snæbjarnarson, V., Steingrímsson, S., & Loftsson, H. (2020). Experimenting with Different Machine Translation Models in Medium-Resource Settings. In P. Sojka, I. Kopeček, K. Pala, & A. Horák (Eds.), *Text, Speech, and Dialogue* (pp. 95–103). Springer. https://doi.org/10.1007/978-3-030-58323-1_10

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Murray, K., Nagata, M., Popel, M., Popovic, M., Shmatova, M., … Zouhar, V. (2024). Preliminary WMT24 Ranking of General MT Systems and LLMs. *arXiv preprint*. https://doi.org/10.48550/arXiv.2407.19884

Kocmi, T., Avramidis, E., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Freitag, M., Gowda, T., Grundkiewicz, R., Haddow, B., Karpinska, M., Koehn, P., Marie, B., Monz, C., Murray, K., Nagata, M., Popel, M., Popović, M., … Zouhar, V. (2024). Findings of the WMT24 General Machine Translation Shared Task: The LLM Era Is Here but MT Is Not Solved Yet. In B. Haddow, T. Kocmi, P. Koehn, & C. Monz (Eds.), *Proceedings of the Ninth Conference on Machine Translation* (pp. 1–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wmt-1.1

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In P. Isabelle, E. Charniak, & D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Association for Computational Linguistics. https://www.aclweb.org/anthology/P02-1040

Popović, M. (2017). chrF++: words helping character n-grams. In O. Bojar, C. Buck, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, & J. Kreutzer (Eds.), *Proceedings of the Second Conference on Machine Translation* (pp. 612–618). Association for Computational Linguistics. https://www.aclweb.org/anthology/W17-4770

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers* (pp. 186–191). Association for Computational Linguistics. https://aclanthology.org/W18-6319

Rei, R., Stewart, C., Farinha, A. C., & Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on*

Empirical Methods in Natural Language Processing (EMNLP) (pp. 2685–2702). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.213

Stap, D., Hasler, E., Byrne, B., Monz, C., & Tran, K. (2024). The Fine-Tuning Paradox: Boosting Translation Quality Without Sacrificing LLM Abilities. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6189–6206). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.336

Tang, K. (2022). PETCI: A Parallel English Translation Dataset of Chinese Idioms. *arXiv preprint*. https://arxiv.org/abs/2202.09509

Úlfarsdóttir, Þ. (2014). ISLEX – a Multilingual Web Dictionary. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2820–2825). European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/pdf/672_Paper.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 5999–6009). http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf