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

Steinþór Steingrímsson, Einar Freyr Sigurðsson, Björn Halldórsson
The Árni Magnússon Institute for Icelandic Studies
{steinst,einasig,bjornh}@hi.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; and (ii) evaluate – using both automatic and manual approaches – three MT systems' abilities to translate MWEs from Icelandic to English. We find that the MT systems evaluated commonly fail when translating idiomatic expressions.

1 Introduction

Multiword expressions (MWEs) are a frequent phenomenon in natural language and speech[1]. 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[2], along with their translations into English as well as structured information about their usage. We classify the MWEs in our dataset *idiomatic expressions*, i.e. idioms with an intended meaning that diverges from the literal meaning of the words constituting the expression, and therefore 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 three machine translation (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.

2 Collecting the Multiword Expressions

The set of multiword expressions, distributed on the Icelandic CLARIN repository[3], 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

[1]We thank three anonymous reviewers for valuable comments on the paper.
[2]http://hdl.handle.net/20.500.12537/275
[3]https://repository.clarin.is/repository/xmlui/handle/20.500.12537/275

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). The first four categories contain type information, cf. for example, the idiom *rétta e-m 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 (/), 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 (nowadays) 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[4].

3 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. It is therefore important to be able to test that ability. To this end, we run three evaluation experiments. First, we simply evaluate the MT output using traditional automatic approaches. We apply the common evaluation metrics BLEU and chrF++. 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]https://github.com/stofnun-arna-magnussonar/IdiomaticExpressions

MT System	BLEU	ChrF	AutoIE (%)
Steingrímsson (filtered parallel data)	9.5	33.2	9.2
Miðeind (using backtranslations)	10.7	33.9	11.5
Google Translate	21.0	49.4	15.5

Table 1: Automatic evaluation of the three MT systems.

3.1 MT Systems

We compare three MT systems that translate from Icelandic into English. The first model is an Icelandic–English translation model (Símonarson et al., 2022) trained by the language-technology startup Miðeind, based on the mBART25 model (Liu et al., 2020). The Miðeind-model is trained on long-context texts, both authentic parallel texts and synthetic texts. The synthetic data comprise backtranslations from various sources, totalling over 150 million tokens. The second model is also based on mBART25. It is only trained on authentic parallel texts, and uses the set of training sentences compiled by Steinþór Steingrímsson using the most effective filtering approach reported in his thesis (Steingrímsson, 2023). The bulk of the training data is from the 21.10 version of the ParIce corpus (Barkarson & Steingrímsson, 2019; Steingrímsson & Barkarson, 2021). Finally, we used one online system, Google Translate[5]. Google Translate was chosen as it is the most popular MT system used by the general public in Iceland.

3.2 Automatic Evaluation Approaches

In order to make a general comparison of the MT systems used, we calculated BLEU and chrF++ for translations of all sentences in the dataset. The scores for BLEU[6] and chrF++[7] were calculated using Sacrebleu (Post, 2018). Sacrebleu signatures are given in footnotes and results reported in Table 1.

Google Translate scored the highest by far, and while Miðeind's system scored slightly higher than Steingrímsson's system, the difference was not statistically significant for the chrF++ scores, as calculated using the pairwise bootstrap test (Koehn, 2004).

Furthermore, we devised a simple automatic approach to gauge how well the MT systems managed to process the idiomatic expressions. Each machine-translated output is either 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. The results of this approach are reported in Table 1, titled AutoIE for Automatic Idiom Evaluation. The score is given as a percentage, representing the ratio of the translated output that the approach classifies as correct.

3.3 Manual Evaluation

To assess whether our automatic approach is useful, all translations of the approx. 1,000 sentences, across all three MT systems were evaluated by a professional translator 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*. He was only to look at the MWE and disregard all other possible errors in the translation. The results are given in Table 2.

Upon inspecting the results, 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

[5]Google Translate was used to translate the sentences on April 22, 2024.
[6]BLEU|nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1
[7]chrF2|nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.3.1

MT System	Correct (%)	Understandable (%)	Incorrect (%)
Steingrímsson (filtered parallel data)	18.8	12.0	69.2
Miðeind (using backtranslations)	25.7	12.5	61.8
Google Translate	37.4	11.3	51.3

Table 2: Manual evaluation of the three MT systems.

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'.

4 Future Work

In most cases, each Icelandic example is given only one translation in our dataset, although more translations may be valid. Adding additional valid translations for each example would be useful in order to use the dataset to automatically evaluate the capabilities of an MT system to translate idiomatic expressions. By comparing the translations deemed correct in the human evaluation to the translations given in the data set, we can add more valid translations. We intend to do so in order to make the data set even more viable for automatic evaluation.

Finally, we intend to use the 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.

5 Conclusions

The evaluation results, and the result analysis, indicate that available MT systems commonly fail when translating idiomatic expressions. Specialized evaluation sets, such as the one introduced in this paper, can be used to gauge the capabilities of MT 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.

References

Barkarson, S., & Steingrímsson, S. (2019). Compiling and Filtering ParIce: An English-Icelandic Parallel Corpus. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 140–145.

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. *Proceedings of the Third International Conference on Language Resources and Evaluation*, 1934–1940.

Koehn, P. (2004). Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 388–395.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., & Zettlemoyer, L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, *8*, 726–742.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318.

Popović, M. (2017). chrF++: words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, 612–618.

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191.

Símonarson, H. B., Jónsson, H. P., Ragnarsson, P. O., Ingólfsdóttir, S. L., Þorsteinsson, V., & Snæbjarnarson, V. (2022). Long Context Translation Models for English-Icelandic translations (22.09) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/278

Steingrímsson, S. (2023). *Effectively compiling parallel corpora for machine translation in resource-scarce conditions* [Doctoral dissertation, Reykjavik University].

Steingrímsson, S., & Barkarson, S. (2021). ParIce: English-Icelandic parallel corpus (21.10) [CLARIN-IS]. http://hdl.handle.net/20.500.12537/145

Úlfarsdóttir, Þ. (2014). ISLEX – a Multilingual Web Dictionary. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 2820–2825.