Examining the Tip of the Iceberg: A Data Set for Idiom Translation

Marzieh Fadaee[1], Arianna Bisazza[2], Christof Monz[1]
[1]Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[2]Leiden Institute of Advanced Computer Science, Leiden University, 2333 CA Leiden, The Netherlands
{m.fadaee,c.monz}@uva.nl
a.bisazza@liacs.leidenuniv.nl

Abstract

Neural Machine Translation (NMT) has been widely used in recent years with significant improvements for many language pairs. Although state-of-the-art NMT systems are generating progressively better translations, idiom translation remains one of the open challenges in this field. Idioms, a category of multiword expressions, are an interesting language phenomenon where the overall meaning of the expression cannot be composed from the meanings of its parts. A first important challenge is the lack of dedicated data sets for learning and evaluating idiom translation. In this paper we address this problem by creating the first large-scale data set for idiom translation. Our data set is automatically extracted from a widely used German↔English translation corpus and includes, for each language direction, a targeted evaluation set where all sentences contain idioms and a regular training corpus where sentences including idioms are marked. We release this data set and use it to perform preliminary NMT experiments as the first step towards better idiom translation.

Keywords: multiword expression, idioms, bilingual corpora, machine translation

1. Introduction

Neural Machine Translation (NMT) (Bahdanau et al., 2015; Sutskever et al., 2014; Cho et al., 2014) has achieved substantial improvements in translation quality over traditional Rule-based and Phrase-based Translation (PBMT) in recent years. For instance, subject-verb agreement, double-object verbs, and overlapping subcategorization are various areas where NMT successfully overcomes the limitations of PBMT (Isabelle et al., 2017; Bentivogli et al., 2016). However, one of the remaining challenges of NMT is translating infrequent words and phrases (Koehn and Knowles, 2017; Fadaee et al., 2017) and idioms are a particular instance of this problem (Isabelle et al., 2017).

Idioms are semantic lexical units whose meaning is often not simply a function of the meaning of its constituent parts (Nunberg et al., 1994; Kövecses and Szabo, 1996). The non-compositionality characteristic of idiom expressions exists in different degrees in a language (Nunberg et al., 1994). In English for example, for the idiom *"spill the beans"*, the word *'spill'* symbolizes *'reveal'* and *'beans'* symbolizes the *'secrets'*. With the idiomatic expression *"kick the bucket"*, on the other hand, no such analysis is possible.

Isabelle et al. (2017) builds a challenge set of 108 short sentences that each focus on one particular difficult phenomenon of the language. Their manual assessment of the eight sentences consisting of an idiomatic phrase show that NMT systems struggle with the translation of these phrases.

The challenge of translating idiom phrases in NMT is partly due to the underlying complexity of identifying a phrase as idiomatic and generating its correct non-literal translation, and partly to the fact that idioms are rarely encountered in the standard data sets used for training NMT systems.

As an example, in Table 1 we provide an idiom expression in German and the literal and idiomatic translations in English. We observe that the literal translation of an idiom is not the correct translation, neither does it capture part of the meaning.

German phrase	*eine weiße Weste haben*
Literal translation	to have a white vest
Idiomatic translation	to have clean slate
Sentence	Coca-Cola und Nestlé gehören zu den Unterzeichnern. Beide **haben** nicht gerade **eine weiße Weste**.
Reference translation	Coca Cola and Nestlé are two signatories with *"spotty" track records*.
DeepL	Coca-Cola and Nestlé are among the signatories. Neither of them is **exactly the same**.
GoogleNMT	Coca-Cola and Nestlé are among the signatories. Both do not **have just a white vest**.
OpenNMT	Coca-Cola and Nestlé are among the signatories. Both don't **have a white essence**.

Table 1: Example of an idiom phrase in German and its translation. We compare the output of DeepL, GoogleNMT, and OpenNMT translating a sentence with this idiom phrase and notice that none capture the idiom translation correctly.

To illustrate the problem of idiom translation we also provide the output of three NMT systems for this sentence: GoogleNMT (Wu et al., 2016), DeepL[1], and the OpenNMT implementation (Klein et al., 2017) based on Bahdanau et al. (2015) and Luong et al. (2015). All systems fail to generate the proper translation of the idiom expression. This problem is particularly pronounced when the source idiom is very different from its equivalent in the target language, as the case here.

[1]www.deepl.com/translator

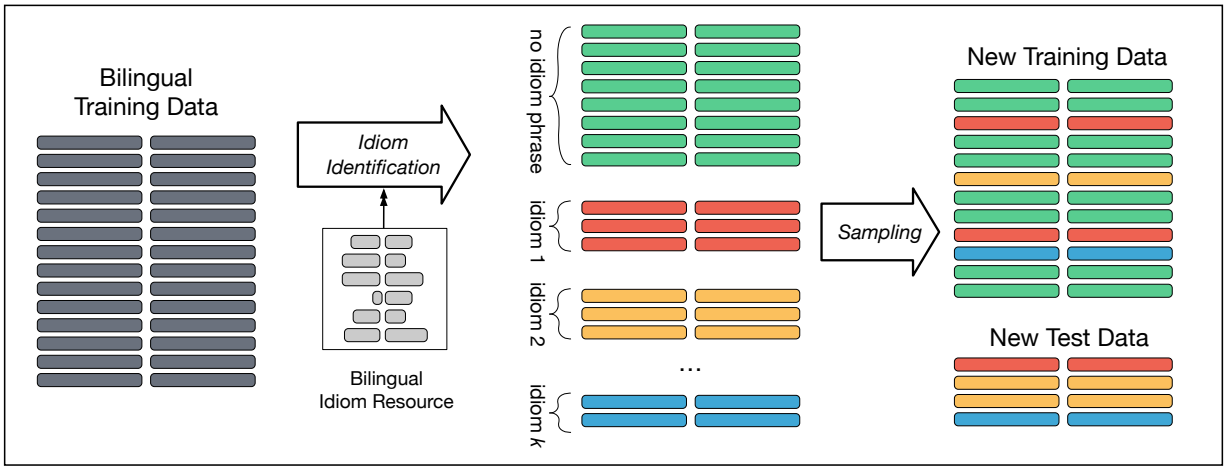

Figure 1: The process of data collection and construction of the test set containing only sentence pairs with idiom phrases.

Although there are a number of monolingual data sets available for identifying idiom expressions (Muzny and Zettlemoyer, 2013; Markantonatou et al., 2017), there is limited work on building a parallel corpus annotated with idioms, which is necessary to investigate this problem more systematically. Salton et al. (2014) selected a small subset of 17 English idioms, collected 10 sentence examples for each idiom from the internet, and manually translated them into Brazilian-Portuguese to use for the translation task.

Building a hand-crafted data set for idiom translation is costly and time-consuming. In this paper we automatically build a new bilingual data set for idiom translation extracted from an existing general-purpose German↔English parallel corpus.

The first part of our data set consists of 1,500 parallel sentences whose German side contains an idiom, while the second consists of 1,500 parallel sentences whose English side contains an idiom. Additionally, we provide the corresponding training data sets for German→English and English→German translation where source sentences including an idiom phrase are marked. We believe that having a sizable data set for training and evaluation is the first step to improve idiom translation.

German idiom translation data set	
Number of unique idioms	103
Training size	4.5M
Idiomatic sentences in training data	1848
Test size	1500
English idiom translation data set	
Number of unique idioms	132
Training size	4.5M
Idiomatic sentences in training data	1998
Test size	1500

Table 2: Statistics of the German and English idiom translation data sets. Sentence pairs are counted on the training and test sets.

2. Data Collection

In this work we focus on German↔English translation of idioms. This is an established language pair and is commonly used in the machine translation community.
Automatically identifying idiom phrases in a parallel corpus requires a gold standard data set annotated manually by linguists. We use the `dict.cc` online dictionary[2] containing idiomatic and colloquial phrases, which is built manually, as our gold standard for extracting idiom phrase pairs. Examining the WMT German↔English test sets from 2008 to 2016 (Bojar et al., 2017), we observe very few sentence pairs containing an idiomatic expression. The standard parallel corpora available for training however contain several such sentence pairs. Therefore we automatically select sentence pairs from the training corpora where the source sentence contains an idiom phrase to build the new test set.
Note that we only focus on idioms on the source side and we have two separate list of idioms for German and English, hence, we independently build two test sets (for German idiom translation and English idiom translation) with different sentence pairs selected from the parallel corpora.

German idiom	*alles über einen kamm scheren*
English equivalent	to measure everything by the same yardstick
Matching German sentence	Aber man kann eben nicht **alle** Inseln **über einen Kamm scheren**.
English translation	But we cannot measure everyone by the same standards.
German idiom	*in den kinderschuhen stecken*
English equivalent	to be in the fledgling stage
Matching German sentence	Es **steckt** immer noch **in den Kinderschuhen**.
English translation	It is still in its infancy.

Table 3: Two examples displaying different constraints of matching an idiom phrase with occurrences in the sentence.

Depending on the language, the words making up an idiomatic phrase are not always contiguous in the sentence.

[2] `www.dict.cc`

German idiom	*in den kinderschuhen stecken*
English equivalent	to be in the fledgling stage
German sentence	Eine Bemerkung, Gentoo/FreeBSD **steckt** noch **in den Kinderschuhen** und ist kein auf Sicherheit achtendes System.
English sentence	Note that Gentoo/FreeBSD is still **in its infancy** and is not a security supported platform.
German idiom	*den kreis schließen*
English equivalent	to bring sth. full circle
German sentence	Die europäische Krise **schließt den Kreis**.
English sentence	The European crisis is **coming full circle**.
German idiom	*auf biegen und brechen*
English equivalent	by hook or crook
German sentence	Nehmen wir zum Beispiel die Währungsunion: Sie soll **auf Biegen und Brechen** eingeführt werden.
English sentence	Take, for example, the introduction -**come what may**- of the single currency.
German idiom	*sie haben das wort*
English equivalent	the floor is yours
German sentence	Berichterstatterin. - (FR) Herr Präsident! Danke, dass **Sie** mir **das Wort** erteilt **haben**.
English sentence	rapporteur. - (FR) Mr President, thank you for **giving** me **the floor**.

Table 4: Examples from the German idiom translation test set.

For instance, in German, the subject can appear between the verb and the prepositional phrase making up the idiom. German also allows for several re-orderings of the phrase.

In order to generalize the process of identifying idiom occurrences, we lemmatize the phrases and consider different re-ordering of the words in the phrase as an acceptable match. We also allow for a fixed number of words to occur in between the words of an idiomatic phrase. Table 3 shows two examples of idiom occurrences that match these criteria.

Following this set of rules, we extract sentence pairs containing idiomatic phrases, and create a set of sentence pairs for each unique idiom phrase. In the next step we sample without replacement from these sets and select individual sentence pairs to build the test set.

In order to build the new training data, we use the remaining sentence pairs in each idiom set as well as the sentence pairs from the original parallel corpora that did not include any idiom phrases. In this process, we ensure that for each idiomatic expression there is at least one occurrence in both training and test data, and that no sentence is included in both training and test data.

Figure 1 visualizes the process of constructing the new training and test sets. As a result, for each language direction, we obtain a targeted test set for idiom translation and the corresponding training corpus representing a natural distribution of sentences with and without idioms.

We annotate each sentence pair with the canonical form of its source-side idiom phrase and its equivalent in the target language.

Table 2 provides some statistics of the two data sets. For each unique idiom in the test set, we also provide the frequency of the respective idiom in the training data. Note that this is based on the lemmatized idiom phrase under the constraints mentioned in Section 2. and is not necessarily an exact match of the phrase.

Table 4 shows several examples from the data set for German idiom translation. We observe that for some idioms the literal translation in the target language is close to the actual meaning, while for others it is not the case.

One side effect of automatically identifying idiom expressions in sentences is that it is not always accurate. Sentence pairs where an idiom expression was used as a literal phrase (e.g., *"spill the beans"* to literally describe the act of *spilling the beans*) will be identified as idiomatic sentences.

3. Translation Experiments

While the main focus of this work is to generate data sets for training and evaluating idiom translation, we also perform a number of preliminary NMT experiments using our data set to measure the problem of idiom translation on large scale data.

In the first experiment following the conventional settings, we do not use any labels in the data to train the translation model. In the second experiment we use the labels in the training data as an additional feature to investigate the effect of informing the model of the existence of an idiomatic phrase in a sentence during training.

We perform a German→English experiment by providing the model with additional input features. The additional features indicate whether a source sentence contains an idiom and are implemented as a special extra token `<idm>` that is prepended to each source sentence containing an idiom. This a simple approach that can be applied to any sequence-to-sequence architecture.

Most NMT systems have a sequence-to-sequence architecture where an encoder builds up a representation of the source sentence and a decoder, using the previous LSTM hidden states and an attention mechanism, generates the target translation (Bahdanau et al., 2015; Sutskever et al., 2014; Cho et al., 2014). We use a 4-layer attention-based encoder-decoder model as described in (Luong et al., 2015) trained with hidden dimension size of 1,000, and batch size of 80 for 20 epochs.

	WMT test sets 2008-2016	Idiom test set		
Model	BLEU	BLEU	Unigram Precision	Word-level Accuracy
PBMT Baseline	20.2	19.7	57.7	71.6
NMT Baseline	26.9	24.8	53.2	67.8
NMT <idm> token on source	25.2	22.5	64.1	73.2

Table 5: Translation performance on German idiom translation test set. *Word-level Idiom Accuracy* and *Unigram Precision* are computed only on the idiom phrase and its corresponding translation in the sentence.

In all experiments the NMT vocabulary is limited to the most common 30K words in both languages and we preprocess source and target language data with Byte-pair encoding (BPE) (Sennrich et al., 2016) using 30K merge operations.

We also use a Phrase-based translation system similar to Moses (Koehn et al., 2007) as baseline to explore PBMT performance for idiom translation.

4. Idiom Translation Evaluation

Ideally idiom translation should be evaluated manually, but this is a very costly process. Automatic metrics, on the other hand, can be used on large data sets at no cost and have the advantage of replicability.

We use the following metrics to evaluate the translation quality with a specific focus on idiom translation accuracy:

BLEU The traditional BLEU score (Papineni et al., 2002) is a good measure to determine the overall quality of the translations. However this measure considers the precision of all n-grams in a sentence and by itself does not focus on the translation quality of the idiomatic expressions.

Modified Unigram Precision To specifically concentrate on the quality of the translation of idiom expressions, we also look at the *localized* precision. In this approach we translate the idiomatic expression in the context of a sentence, and only evaluate the translation quality of the idiom phrase.

To isolate the idiom translation in the sentence, we look at the word-level alignments between the idiom expression in the source sentence and the generated translation in the target sentence. We use `fast-align` (Dyer et al., 2013) to extract word alignments. Since idiomatic phrases and the respective translations are not contiguous in many cases we only compare the unigrams of the two phrases.

Note that for this metric we have two references: The idiom translation as an independent expression, and the human generated idiom translation in the target sentence.

Word-level Idiom Accuracy We also use another metric to evaluate the word-level translation accuracy of the idiom phrase. We use word alignments between source and target sentences to determine the number of correctly translated words. We use the following equation to compute the accuracy:

$$WIAcc = \frac{H - I}{N}$$

where H is the number of correctly translated words, I is the number of extra words in the idiom translation, and N is the number of words in the gold idiom expression.

Table 5 presents the results for the translation task using different metrics. Looking at the overall BLEU scores, we observe that baseline performance on the idiom-specific test set is lower than on the union of the standard test sets (WMT 2008-2016). While the scores on these two data sets are not directly comparable, this result is in line with previous findings that sentences containing idiomatic expressions are harder to translate (Isabelle et al., 2017). We can also see that the performance gap is not as pronounced for PBMT systems, suggesting that phrase-based models are capable of *memorizing* the idiom phrases to some extent.

The NMT experiment using a special input token to indicate the presence of an idiom in the sentence performs still better than PBMT but slightly worse than the NMT baseline in terms of BLEU. Despite this drop in BLEU performance, by examining the *unigram precision* and *word-level idiom accuracy* scores, we observe that this model generates more accurate idiom translations.

These preliminary experiments reiterate the problem of idiom translation with neural models, and in addition show that with a labeled data set, we can devise simple models to address this problem to some extent.

5. Conclusion

Idiom translation is one of the more difficult challenges of machine translation. Neural MT in particular has been shown to perform poorly on idiom translation despite its overall strong advantage over previous MT paradigms (Isabelle et al., 2017). As a first step towards a better understanding of this problem, we have presented a parallel data set for training and testing idiom translation for German→English and English→German.

The test sets include sentences with at least one idiom on the source side while the training data is a mixture of idiomatic and non-idiomatic sentences with labels to distinguish between the two. We also performed preliminary translation experiments and proposed different metrics to evaluate idiom translation.

We release new data sets which can be used to further investigate and improve NMT performance in idiom translation.

Acknowledgments

This research was funded in part by the Netherlands Organization for Scientific Research (NWO) under project numbers 639.022.213 and 639.021.646, and a Google Faculty Research Award. We also thank NVIDIA for their hardware support.

6. Bibliographical References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus phrase-based machine translation quality: a case study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, Texas, November. Association for Computational Linguistics.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. (2017). Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 169–214, Copenhagen, Denmark, September. Association for Computational Linguistics.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada, July. Association for Computational Linguistics.

Isabelle, P., Cherry, C., and Foster, G. (2017). A challenge set approach to evaluating machine translation. *arXiv preprint arXiv:1704.07431*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Kövecses, Z. and Szabo, P. (1996). Idioms: A view from cognitive semantics. *Applied Linguistics*, 17(3):326–355.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

Markantonatou, S., Ramisch, C., Savary, A., and Vincze, V. (2017). Proceedings of the 13th workshop on multiword expressions (mwe 2017). In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*.

Muzny, G. and Zettlemoyer, L. (2013). Automatic idiom identification in wiktionary. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1417–1421.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Salton, G., Ross, R., and Kelleher, J. (2014). 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, April. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August. Association for Computational Linguistics.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, et al., editors, *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.