Investigating Terminology Translation in Statistical and Neural Machine Translation: A Case Study on English-to-Hindi and Hindi-to-English

Rejwanul Haque,[†] **Mohammed Hasanuzzaman**[‡] **and Andy Way**[†]
ADAPT Centre
[†]School of Computing, Dublin City University, Dublin, Ireland
[‡]Department of Computer Science, Cork Institute of Technology, Cork, Ireland
firstname.lastname@adaptcentre.ie

Abstract

Terminology translation plays a critical role in domain-specific machine translation (MT). In this paper, we conduct a comparative qualitative evaluation on terminology translation in phrase-based statistical MT (PB-SMT) and neural MT (NMT) in two translation directions: English-to-Hindi and Hindi-to-English. For this, we select a test set from a legal domain corpus and create a gold standard for evaluating terminology translation in MT. We also propose an error typology taking the terminology translation errors into consideration. We evaluate the MT systems' performance on terminology translation, and demonstrate our findings, unraveling strengths, weaknesses, and similarities of PB-SMT and NMT in the area of term translation.

1 Introduction

Over the last five years, there has been incremental progress in the field of NMT (Bahdanau et al., 2015; Vaswani et al., 2017) to the point where some researchers are claiming parity with human translation (Hassan et al., 2018). Nowadays, NMT is regarded as a preferred alternative to PB-SMT (Koehn et al., 2003) and represents a new state-of-the-art in MT research. The rise of NMT has resulted in a swathe of research in the field of MT, unraveling the strengths, weaknesses, impacts and commercialisation aspects of the classical (i.e. PB-SMT) and emerging (i.e. NMT) methods (e.g. (Bentivogli et al., 2016; Toral and Way, 2018)). In brief, the NMT systems are often able to produce better translations than the PB-SMT systems. Interestingly, terminology translation, a crucial factor in industrial translation workflows (TWs), is one of the less explored areas in MT research. In this context, a few studies (Burchardt et al., 2017; Macketanz et al., 2017; Specia et al., 2017), with their focus on high-level evaluation, have indicated that NMT lacks effectiveness in translating domain terms compared to PB-SMT. In this work, we aim to compare PB-SMT and NMT in relation to terminology translation, by carrying out a thorough manual evaluation. For this, we select a test set from legal domain data (i.e. judicial proceedings), and create a gold standard evaluation test set following a semi-automatic terminology annotation strategy. We inspected the patterns of the term translation-related errors in MT. From our observations we make a high-level classification of the terminology translation-related errors and propose an error typology. We discuss various aspects of terminology translation in MT considering each of the types from the proposed terminology translation typology, and dig into the extent of the term translation problems in PB-SMT and NMT with statistical measures as well as linguistic analysis. For experimentation, we select a less examined and low-resource language pair, English–Hindi.

2 MT Systems

To build our PB-SMT systems we used the Moses toolkit (Koehn et al., 2007). For LM training we combine a large monolingual corpus with the target-side of the parallel training corpus. Additionally, we trained a neural LM with the NPLM toolkit (Vaswani et al., 2013) on the target side of the parallel training corpus alone. We considered the standard PB-SMT log-linear features for training. We call the English-to-Hindi and Hindi-to-English PB-SMT systems EHPS and HEPS, respectively. Our NMT systems are Google Transformer models (Vaswani et al., 2017). In our experiments we followed the recommended best set-up from Vaswani et al. (2017). We call our the English-to-Hindi and Hindi-to-English NMT systems EHNS and HENS, respectively.

Proceedings of Recent Advances in Natural Language Processing, pages 437–446,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_052

For experimentation we used the IIT Bombay English-Hindi parallel corpus (Kunchukuttan et al., 2017). For building additional LMs for Hindi and English we use the HindEnCorp monolingual corpus (Bojar et al., 2014) and monolingual data from the OPUS project (Tiedemann, 2012), respectively. Corpus statistics are shown in Table 1. We selected 2,000 sentences (test set) for the evaluation of the MT systems and 996 sentences (development set) for validation from the Judicial parallel corpus (cf. Table 1) which is a juridical domain corpus (i.e. proceedings of legal judgments). The MT systems were built with the training set shown in Table 1 that includes the remaining sentences of the Judicial parallel corpus.

Table 1: Corpus Statistics.

English–Hindi parallel corpus			
	Sentences	Words (En)	Words (Hi)
Training set	1,243,024	17,485,320	18,744,496
(Vocabulary)		180,807	309,879
Judicial	7,374	179,503	193,729
Development set	996	19,868	20,634
Test set	2,000	39,627	41,249
Monolingual Corpus		Sentences	Words
Used for PB-SMT Language Model			
English		11M	222M
Hindi		10.4M	199M
Used for NMT Back Translation			
English		1M	20.2M
Hindi		903K	14.2M

We present the comparative performance of the PB-SMT and NMT systems in terms of BLEU score (Papineni et al., 2002) in Table 2. Additionally, we performed statistical significance tests using bootstrap resampling methods (Koehn, 2004). The confidence level (%) of the improvement obtained by one MT system with respect to the another MT system is reported. As can be seen from Table 2, EHPS and EHNS produce reasonable BLEU scores (28.8 BLEU and 36.6 BLEU) on the test set given the difficulty of the translation pair. These BLEU scores, in fact, underestimate the translation quality, given the relatively free word order in Hindi, as we have just a single reference translation set for evaluation. As far as the Hindi-to-English translation task is concerned, HEPS and HENS produce moderate BLEU scores (34.1 BLEU and 39.9 BLEU) on the test set. As expected, translation quality in the morphologically-rich to morphologically-poor language improves.

Table 2: Performance of MT systems on BLEU.

System	BLEU	System	BLEU
EHPS	28.8	HEPS	34.1
EHNS	**36.6** (99.9%)	HENS	**39.9** (99.9%)

3 Creating Gold Standard Evaluation Set

To evaluate terminology translation with our MT systems, we manually annotated the test set by marking term-pairs on the source- and target-sides of the test set (cf. Table 1) with a view to creating a gold standard evaluation set. The annotation process is performed using our own bilingual term annotation tool, *TermMarker*. If there is a source term present in the source sentence, its translation equivalent (i.e. target term) is found in the target sentence, and the source–target term-pair is marked. The annotators are native Hindi evaluators with excellent English skills. They were instructed to mark those words as terms that belong to legal or judicial domains. The annotators were also instructed to mark those sentence-pairs from the test set that contain errors (e.g. mistranslations, spelling mistakes) in either source or target sentences. The annotators reported 75 erroneous sentence-pairs which we discarded from the test set. In addition, 655 sentence-pairs of the test set did not contain any terms. We call the remaining 1,270 sentence-pairs our *gold-testset*. Each sentence-pair of gold-testset contains at least one aligned source-target term-pair. We have made the gold-testset publicly available to the research community.[1]

Annotation Suggestions from Bilingual Terminology While manually annotating bilingual terms in the judicial domain test set, we took support from a bilingual terminology that was automatically created from the Judicial corpus (cf. Table 1). For automatic bilingual term extraction we followed the approach of Haque et al. (2018). We found 3,064 English terms and their target equivalents (3,064 Hindi terms) in the source- and target-sides of gold-testset, respectively.

Variations of Term A term may have more than one domain-specific translation equivalent. The number of translation equivalents for a source term could vary from language to language depending on the morphological nature of the target language. For example, translation of the En-

[1] https://www.computing.dcu.ie/~rhaque/termdata/terminology-testset.zip

glish word ‘affidavit’ has multiple target equivalents (LIVs (lexical and inflectional variations)) in Hindi even if the translation domain is legal or juridical: ‘shapath patr’, ‘halaphanaama’, ‘halaphanaame’, or ‘halaphanaamo’. The term ‘shapath patr’ is the lexical variation of Hindi term ‘halaphanaama’. The base form ‘halaphanaama’ could have many inflectional variations (e.g. ‘halaphanaame’, ‘halaphanaamo’) given the sentence’s syntactic and morphological profile (e.g. gender, case).

For each term we check whether the term has any additional LIVs pertaining to the juridical domain and relevant to the context of the sentence. If this is the case, we include the relevant variations as legitimate alternatives term.

We again exploit the method of Haque et al. (2018) for obtaining variation suggestions for a term. The automatically extracted bilingual terminology of Haque et al. (2018) comes with the four highest-weighted target terms for a source term. If the annotator accepts an annotation suggestion (source–target term-pair) from the bilingual terminology, the remaining three target terms are considered as alternative suggestions of the target term.

Two annotators took part in the annotation task, and two sets of annotated data were obtained. The term-pairs of gold-testset are finalised on the basis of the annotation agreement by the two annotators, i.e. we keep those source–target term-pairs in gold-testset for which both annotators agree that the source and target entities are terms and aligned. On completion of the annotation process, inter-annotator agreement was computed using Cohen’s kappa (Cohen, 1960) at word-level. For each word we count an agreement whenever both annotators agree that it is a term (or part of term) or non-term entity. We found the kappa coefficient to be very high (i.e. 0.95) for the annotation task. This indicates that our terminology annotation is of excellent quality.

The final LIV list for a term is the union of the LIV lists created by the annotators. This helps make the resulting LIV lists exhaustive.

4 Terminology Translation Typology

In order to annotate errors in (automatic) translations, MT users often exploit the MQM (Multidimensional Quality Metric) error annotation framework (Lommel et al., 2014). One of the error types in the MQM toolkit is terminology (i.e. *inconsistent with termbase, inconsistent use of terminology*) which is an oversimplified attribute and does not consider various nuances of term translation errors. We propose an error typology taking terminology translation into consideration. First, we translated the test set sentences with our MT systems, and sampled 300 translations from the whole translation set. Then, the terminology translations were manually inspected, noting the patterns of the term translation-related errors. From our observations we found that the terminology translation-related errors can be classified into eight primary categories. As far as the term translation quality of an MT system is concerned, our proposed typology could provide a better perspective as to how the MT system lacks quality in translating domain terms. The categories are as follows: (i) reorder error (RE): the translation of a source term forms the wrong word order in the target, (ii) inflectional error (IE): the translation of a source term inflicts a morphological error, (iii) partial error (PE): the MT system correctly translates part of a source term into the target and commits an error for the remainder of the source term, (iv) incorrect lexical selection (ILS): the translation of a source term is an incorrect lexical choice, (v) term drop (TD): the MT system omits the source term in translation, (vi) source term copied (STC): a source term or part of it is copied verbatim to target, (vii) disambiguation issue in target (DIT): although the MT system makes a potentially correct lexical choice for a source term, its translation-equivalent does not carry the meaning of the source term, and (viii) other error (OE): there is an error in relation to the translation of a source term, whose category, however, is beyond all remaining error categories. The proposed terminology translation error typology is illustrated in Figure 1 (cf. Appendix A).

Apart from the above error categories, we have a class for a source term being correctly translated into the target, i.e. the MT system produces a correct translation (CT) for a source term. As pointed out in Section 3, we wanted to see how diverse an MT model can be in translating domain terms, and how close the translation of a source term can be to the reference terms or its LIVs or to what extent (e.g. syntactically and morphologically) it differs from them. For this reason, we divide the CT class into seven sub-classes, and define them below: (i) CT given the reference term (CTR): the translation of a source term is the reference term, (ii) CT given one of the LIVs (CTV): the translation of a source is one of the LIVs of the reference term, (iii) variation missing (VM): a source term is correctly translated into the target, but the translation is neither the reference term nor any of its LIVs, (iv) correct inflected form (CIF): a source term is correctly translated into the target, but the translation is neither the reference term nor any of its LIVs. However, the base form of the translation of the source term is identical to the base form of either the reference term or one of the LIVs of the reference term, (v) correct reorder form (CRF): a source term is correctly translated into the target, and the translation includes those words that either the reference term or one of the LIVs has, but the word order of the translation is different to that of the the reference term or one of the LIVs, (vi) correct reorder and inflected form (CRIF): this class is a combination of both CIF and CRF, and (vii) other correct (OC): a source term is correctly translated into the target, whose category, however, is beyond the all remaining correct categories.

5 Manual Evaluation Plan

This section presents our manual evaluation plan. Translations of the source terms of gold-testset were manually validated and classified in accordance with the set of fine-grained errors and correct categories described above. This was accomplished by the human evaluator. The manual evaluation was carried out with a GUI that randomly displays a source sentence and its reference translation from gold-testset, and the automatic translation by one of the MT systems. For each source term the GUI highlights the source term and the corresponding reference term from the source and reference sentences, respectively, and displays the LIVs of the reference term, if any. The GUI lists the error and correct categories described in Section 4. The evaluator, a native Hindi speaker with the excellent English and Hindi skills, was instructed to follow the following criteria for evaluating the translation of a source term: (a) judge correctness / incorrectness of the translation of the source term in hypothesis and label it with an appropriate category listed in the GUI, (b) do not need to judge the whole translation, but instead look at the local context to which both source term and its translation belong, and (c) take the syntactic and morphological properties of the source term and its translation into account.

The manual classification process was completed for all MT system types. We measure agreement in manual classification of terminology translation. For this, we randomly selected an additional 100 segments from gold-testset and hired another evaluator having the similar skills. We considered the correct and incorrect categories for the calculation, i.e. we count an agreement whenever both evaluators agree that it is a correct (or incorrect) term translation, with agreement by chance = 1/2. We found that the kappa coefficient for this ranges from 0.97 to 1.0. Thus, our manual term translation classification quality can be labeled as excellent.

6 Terminology Translation Evaluation in PB-SMT and NMT

This section provides a comparative evaluation of the ability of PB-SMT and NMT to translate terminology accurately. In Table 5, we report the statistics of terminology translations from the English-to-Hindi MT task. We see that EHPS and EHNS incorrectly translate 303 and 253 English terms (out of total 3,064 terms) (cf. last row of Table 5), respectively, into Hindi, resulting in 9.9% and 8.3% terminology translation errors, respectively. We use approximate randomization (Yeh, 2000) to test the statistical significance of the difference between two systems, and report the significance-level (*p*-value) in the last column of Table 5. We found that the difference between the error rates is statistically significant. In Table 6, we report the statistics of terminology translations for the Hindi-to-English MT task. We see that HEPS and HENS incorrectly translate 396 and 353 Hindi terms (cf. last row of Table 6), respectively, into English, resulting in 12.9% and 11.5% terminology translation errors, respectively. As can be seen from Table 6, the difference between the error rates is statistically significant. When we compare these scores with those from Table 5, we see that these scores are slightly higher compared to those for the English-to-Hindi task. Surprisingly, the terminology translation quality from the morphologically-rich to the morphologically-poor language deteriorates compared to the overall MT quality (cf. Section 2).

6.1 Comparison with Fine-Grained Category

This section discusses the numbers and highlights phenomena for the fine-grained categories, starting with those that involve correct terminology translations.

CTV & VM We see from Tables 5 and 6 that the numbers under the CTV (correct term given one of the LIVs class are much higher in the English-to-Hindi task (695 and 662) compared to those in the Hindi-to-English task (241 and 245). CTV is measured as the count of instances where a source term is (i) correctly translated into the target translation and (ii) the translation-equivalent of that term is one of the LIVs of the reference term. As can be seen from Table 1, the training set vocabulary size is much higher in Hindi compared to that in English since the former is a morphologically-rich and highly inflected language, which is probably the reason why these numbers are much higher in the English-to-Hindi task.

In a few cases, the human evaluator found that the source terms are correctly translated into the target, but the translations are neither the reference terms nor any of its LIVs. The manual evaluator marked those instances with VM (variation missing) (cf. Tables 5 and 6). These can be viewed as annotation mistakes since the annotator omitted to add relevant LIVs for the reference term into gold-testset. In future, we aim to make gold-testset as exhaustive as possible by adding missing LIVs for the respective reference terms.

CRF, CIF, CRIF & OC We start this section by highlighting the problem of word order in term translation, via a translation example from gold-testset. The Hindi-to-English NMT system correctly translates a Hindi source term 'khand nyaay peeth ke nirnay' (English reference term: 'division bench judgment') into the following target translation (English): "it shall also be relevant to refer to article 45 - 48 of the *judgment of the division bench*". The manual evalautor marks this term translation as CRF (correct reorder form) since the term *'judgment of the division bench'* was not in the LIV list for the reference term, 'division bench judgment'.

We show another example from the Hindi-to-English translation task. This time, we highlight the issue of inflection in term translation. As an example, we consider a source Hindi term 'abhikathan' from gold-testset. Its reference term is 'allegation', and the LIV list of the reference term includes two lexical variations for 'allegation': 'accusation' and 'complaint'. A portion of the reference translation is 'an allegation made by the respondent ...'. A portion of the translation produced by the Hindi-to-English NMT system is 'it was *alleged* by the respondent ...'. In this translation, we see the Hindi term 'abhikathan' is translated into 'alleged' which is a correct translation of the Hindi legal term 'abhikathan' as per the syntax of the target translation. As above, the manual evalautor marked these term translations as CIF (correct inflected form) since the translation-equivalent of this term is not found in the LIV list of the reference term.

As stated in Section 4, CRIF (correct reorder and inflected form) is the combination of the above two types: CRF and CIF. As an example, consider a portion of the source Hindi sentence '*vivaadagrast vaseeyat* hindee mein taip kee gaee hai ...' and the English reference translation 'the *will in dispute* is typed in hindi ...' from gold-testset. Here, 'vivaadagrast vaseeyat' is a Hindi term and its English equivalent is 'will in dispute'. The translation of the source sentence by the Hindi-to-English NMT system is 'the *disputed will* have been typed in hindi ...'. We see that the translation of the source term ('vivaadagrast vaseeyat') is 'disputed will' which is correct. We also see that its word order is different to that of the reference term ('will in dispute'); and the morphological form of (part of) the translation is not identical to that of (part of) the reference term. As is the case with CRF and CIF, the manual evaluator marks such term translations as CRIF.

When translation of a source term is correct but its category is beyond the all remaining correct categories, the manual evaluator marks that term translation as OC (other correct). In our manual evaluation task, we encountered various such phenomena, and detail some of those below. (1) term transliteration: the translation-equivalent of a source term is the transliteration of the source term itself. We observed this happening only when the target language is Hindi. In practice, many English terms (transliterated form) are often used in Hindi text (e.g. 'decree' as 'dikre', 'tariff orders' as 'tarif ordars'), (2) terminology translation coreferred: translation-equivalent of a source term is not found in the hypothesis, however, it is correctly coreferred in target translation, and (3) semantically coherent terminology translation: the translation-equivalent of a source term is not seen in the hypothesis, but its meaning is correctly transferred into the target. As an example, consider the source Hindi sentence "sabhee apeelakartaon ne *aparaadh sveekaar nahin* kiya aur muqadama chalaaye jaane kee maang kee", and reference English sentence "all the appellants pleaded not guilty to the charge and claimed to be tried" from gold-testset.[2] Here, 'aparaadh sveekaar nahin' is a Hindi term and its English translation is 'pleaded not guilty'. The Hindi-to-English NMT system produces the following English translation "all the appellants did not accept the crime and sought to run the suit" for the source sentence. In this example, we see the meaning of

[2] In this example, the reference English sentence is the literal translation of the source Hindi sentence.

the source term 'aparaadh sveekaar nahin' is preserved in the target translation.

Table 3: CIF, CRF, CRIF and OC in PB-SMT and NMT.

	PB-SMT	NMT
English-to-Hindi	122 (4%)	98 (3.2%)
Hindi-to-English	90 (2.9%)	138 (4.5%)

We recall the rule that we defined while forming the LIV list for a reference term from Section 3. Our annotators considered only those inflectional variations for a reference term that would be grammatically relevant to the context of the reference translation in which they would appear. In practice, translation of a source sentence can be generated in numerous ways. It is possible that a particular inflectional variation of a reference term could be grammatically relevant to the context of the target translation, which, when it replaces the reference term in the reference translation, may (syntactically) misfit the context of the reference translation. As far as CRF and CRIF are concerned, a similar story might be applicable to the translation of a multiword term. A multiword term may be translated into the target in various ways (as shown above, 'division bench judgment' as 'judgment of the division bench', and 'disputed will' as 'will in dispute'). In reality, it would be an impossible task for the human annotator to consider all possible such variations for a multiword reference term. Additionally, as above, we saw more diverse translations with the domain terms under the OC category. In Table 3, we report the combined numbers under the above categories (CRF, CRIF, CIF and OC), with their percentage with respect to the total number of terms. We see that translations of a notable portion of source terms in each translation task are diverse. Therefore, investigating the automation of the terminology translation evaluation process (Haque et al., 2019), these phenomena have to be taken into consideration.

RE Now, we turn our focus to the error classes, starting with RE (reordering error). We compare the results under RE from Tables 5 and 6, and we see that NMT commits many fewer terminology translation-related reordering errors than PB-SMT. 15 REs are caught in the English-to-Hindi PB-SMT task compared to 5 in the English-to-Hindi NMT task. The same trend is observed with the reverse direction, with 18 reordering errors seen in the Hindi-to-English PB-SMT task compared to 5 in the Hindi-to-English NMT task. As can be seen from the last columns of Tables 5 and 6, the differences in these numbers in PB-SMT and NMT are statistically significant.

IE As far as the inflectional error type is concerned, the Hindi-to-English PB-SMT system makes nearly twice as many mistakes as the Hindi-to-English NMT system (118 vs 76) (cf. Tables 5 and 6), which is statistically significant. We see a different picture in the English-to-Hindi direction, i.e. the numbers of morphological errors are nearly the same, both in PB-SMT and NMT (77 vs 79). We found no statistically significant difference between them.

PE The numbers (cf. Tables 5 and 6) of partial term translation errors in PB-SMT and NMT are almost the same regardless of the translation directions. We found that the differences in these numbers are not statistically significant.

ILS PB-SMT appears to be more error-prone than NMT as far as a term's lexical selection is concerned. EHPS commits 77 incorrect lexical choices which is 35 more than EHNS. The same trend is observed with the Hindi-to-English direction. HEPS and HENS commit 139 and 90 incorrect lexical choices, respectively. We found that the differences in these numbers in PB-SMT and NMT are statistically significant.

TD Comparing the numbers of the term drop category from Tables 5 and 6, we see that the numbers of term omission by the PB-SMT and NMT systems are almost the same (53 versus 56) in the English-to-Hindi translation task. We found no statistically significant difference in these numbers. In contrast, in the Hindi-to-English translation task, HENS drops terms more than twice as often as HEPS (86 versus 38). This time, we found that the difference in these numbers is statistically significant.

STC & OE Now we focus on discussing various aspects with the STC (source term copied) and OE (other error) classes, starting with the English-to-Hindi task. We counted the number of source terms of gold-testset that are not found in the source-side of the training corpus (cf. Table 1). We see that 88 source terms (out of a total of 3,064 terms) are not found in the training data, with almost all being multiword terms. Nevertheless, only 5 unique words (i.e. adjudicary, halsbury, presuit, decretal, adj) that are either single-word terms or words of multiword terms are not found in the training data. In other words, these are out-of-vocabulary (OOV) items.

Table 4: STC in English-to-Hindi PB-SMT and NMT.

STC (PB-SMT)	translation (NMT)	class (NMT)
adjudicatory role	nyaay - nirnay keea koee bhoomika	PE
decretal		TD
halsbury 's laws	halbury ke kaanoonon	PE
presuit	poorva vaad	RE
adjudicatory		TD
learned **adj**	vidvat edeeje	CTR
learned **adj**	kaabil edeeje	CTV
mrtp act	mrtp adhiniyam	CTR
testatrix	testrex	OE
concealments	rahasyon	OE
res **judicata**	nyaayik roop	OE
subjudice	vichaaraadheen	CTV

We recall Table 5 where we see that the manual evaluator has marked 12 term translations with STC since in those cases the PB-SMT system copied source terms (or a part of source terms) verbatim into the target. In Table 4, we show those source terms in the PB-SMT task that belong to the STC class. The first column of the table shows source terms with the term itself or part of it in bold, which means those words are copied verbatim into target. We see from the table that the OOV terms (i.e. adjudicary, halsbury, presuit, decretal, adj), in most cases, are responsible for the term translations being marked with the STC tag. In one instance we found that a part of the English term ('mrtp') (cf. row 8 of Table 4) itself was present in the target-side of the training corpus. This could be the possible reason why 'mrtp' is seen in the target translation. Each of the remaining source terms (last 4 rows of Table 4) include words that are copied directly into the target translation despite the fact that they are not OOVs. This is a well-known problem in PB-SMT and rarely happens with the low frequency words of the training corpus. In short, these source terms (last 4 entries of Table 4) either alone or with the adjacent words of the test set sentences (i.e. as a part of phrase) are not found in the source-side of the PB-SMT phrase table.

Now we see how NMT performed with the 12 source terms above; their translations with EHNS and the corresponding manual class are shown in the second and third columns of Table 4, respectively. We see that out of 12 translations EHNS made a mistake on 8 occasions and correctly translated on 4 occasions. The errors are spread over different categories (e.g. TD, OE, PE). Unsurprisingly, we see NMT is capable of correctly translating rare and even unknown words, by exploiting the strength of the open-vocabulary translation technique (Sennrich et al., 2016). However, this method also has down-sides. For example, some of the term translations under the OE category in the NMT task are non-existent wordforms of the target language, for which the open-vocabulary translation technique is responsible. This phenomenon is also corroborated by Farajian et al. (2017) while translating technical domain terms. We discuss the OE class further below.

We see from Table 5 that the human evaluator has marked 24 term translations with OE in NMT. In this category we observed that the translations of the source terms are usually either strange words that have no relation to the meaning of the source term, repetitions of other translated words or terms, entities that are non-existent wordforms of the target language, or words with typographical errors. As far as PB-SMT is concerned, we see from Table 5 that the evaluator also tagged 12 term translations with OE, most of which are related to typographical errors.

Now we turn our focus on the Hindi-to-English task. We counted the number of those source terms from gold-testset that are not found in the source-side (Hindi) of the training corpus (cf. Table 1). We see that 160 source terms (out of a total of 3,064 terms) are not found in the training data, most of which are, in fact, multiword terms. However, only 18 unique Hindi words that are either single-word terms or words within multiword terms are not found in the training data. As in English-to-Hindi translation task, in this task we found that the OOV items are largely responsible for the term translations being marked as STC. We also examined how the Hindi-to-English NMT system performed with those 17 source terms that were marked as STC. We see that HENS makes a mistake on 13 occasions and correctly translates on 4 occasions. The error types are spread over different categories: TD (2), OE (6), PE (1) and ILS (4). We observed that 3 out of 4 source terms of the STC category for which the Hindi-to-English NMT system produces correct translations are OOV items. Here, we again see the strength of the open-vocabulary translation technique for the translation of novel terms. In the Hindi-to-English translation task, we found that the terminology translations under the OE category, as in English-to-Hindi translation, are roughly related to odd translations, non-existent wordforms of the target language, typological mistakes and repetition of other translated words or terms.

DIT We see from Table 5 and Table 6 that the manual evaluator marked 3 and 1 term translations as DIT (disambiguation issue in target) in English-to-Hindi and Hindi-to-English PB-SMT tasks, respectively. We found that the MT systems made correct lexical choices for the source terms, although the meanings of their target-equivalents in the respective translations are different to those of the source terms. This can be viewed as a cross-lingual disambiguation problem. For example, one of the three source terms from English-to-Hindi translation task is 'victim' (reference translation 'shikaar') and the English-to-Hindi PB-SMT system makes a correct lexical choice ('shikaar') for 'victim', although the meaning of 'shikaar' is completely different in the target translation, i.e. here, its meaning is equivalent to English 'hunt'.

Pairwise Overlap We report the numbers of pairwise overlaps, i.e. the number of instances in which NMT and PB-SMT have identical classification outcomes. We recall Table 5 & 6 whose fourth columns show the numbers of pairwise overlap for categories. The small number of overlapping instances in each category indicates that term translation errors from the PB-SMT system are quite different from those from the NMT system. As can be seen from the last row of Table 5 & 6, the numbers of overlaps in the combination of all error classes are 86 and 115, respectively, which are nearly one third or fourth of the number of errors committed by the NMT and PB-SMT systems alone, indicating that the majority of the errors in PB-SMT are complementary with those in NMT. This finding on terminology translation is corroborated by Popović (2017), who finds complementarity with the various issues relating to the translations of NMT and PB-SMT.

7 Conclusion and Future Work

In this paper, we investigated domain term translation in PB-SMT and NMT with two morphologically divergent languages, English and Hindi. Due to the unavailability of a gold standard for term translation evaluation, we adopted a technique that semi-automatically creates a gold standard test set from an English–Hindi judicial domain parallel corpus. The gold standard that we have developed will serve as an important resource for the evaluation of term translation in MT. We also propose a terminology translation typology focused on term translation errors in MT. From our evaluation results, we found that the NMT systems commit fewer lexical, reordering and morphological errors than the PB-SMT systems. The differences in error rates of the former (lexical selection and reordering errors) types are statistically significant in both MT tasks, and the difference of the morphological error rates is statistically significant in the Hindi-to-English task. The morphological errors are seen relatively more often in PB-SMT than in NMT when translation is performed from a morphologically-rich language (Hindi) to the a morphologically-poor language (English). The opposite picture is observed in the case of term omission in translation, with NMT omitting more terms in translation than PB-SMT. We found that the difference in term omission-related error rates in PB-SMT and NMT are statistically significant in the Hindi-to-English task, i.e. again from the morphologically-rich language to the morphologically-poor language. Another important finding from our analysis is that NMT is able to correctly translate unknown terms, by exploiting the strength of the open-vocabulary translation technique, which, as expected, are copied verbatim into the target in PB-SMT. We also found that the majority of the errors made by the PB-SMT system are complementary to those made by the NMT system. In NMT, we observed that translations of source terms are occasionally found to be strange words that have no relation to the source term, non-existent wordforms of the target language, and/or repetition of other translated words. This study also shows that a notable portion of the term translations by the MT systems are diverse, which needs to be taken into consideration while investigating the automation of the terminology translation evaluation process.

As far as future work is concerned, we plan to test terminology translation with different language pairs and domains.

Acknowledgments

The ADAPT Centre for Digital Content Technology is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund. This project has partially received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713567, and the publication has emanated from research supported in part by a research grant from SFI under Grant Number 13/RC/2077.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

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.

Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., and Zeman, D. (2014). HindEnCorp – Hindi-English and Hindi-only corpus for machine translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC'14)*, pages 3550–3555, Reykjavik, Iceland.

Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.-T., and Williams, P. (2017). A linguistic evaluation of rule-based, phrase-based, and neural mt engines. *The Prague Bulletin of Mathematical Linguistics*, 108(1):159–170.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Farajian, M. A., Turchi, M., Negri, M., Bertoldi, N., and Federico, M. (2017). Neural vs. phrase-based machine translation in a multi-domain scenario. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 280–284, Valencia, Spain.

Haque, R., Hasanuzzaman, M., and Way, A. (2019). TermEval: An automatic metric for evaluating terminology translation in MT. In *Proceedings of CICLing 2019, the 20th International Conference on Computational Linguistics and Intelligent Text Processing*, La Rochelle, France.

Haque, R., Penkale, S., and Way, A. (2018). TermFinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Language Resources and Evaluation*, 52(2):365–400.

Hassan, H., Aue, A., Chen, C., Chowdhary, V., Clark, J., Federmann, C., Huang, X., Junczys-Dowmunt, M., Lewis, W., Li, M., Liu, S., Liu, T., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., and Zhou, M. (2018). Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In Lin, D. and Wu, D., editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., College, W., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions*, pages 177–180, Prague, Czech Republic.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, pages 48–54, Edmonton, AB.

Kunchukuttan, A., Mehta, P., and Bhattacharyya, P. (2017). The IIT Bombay English–Hindi parallel corpus. *CoRR*, 1710.02855.

Lommel, A. R., Uszkoreit, H., and Burchardt, A. (2014). Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica: tecnologies de la traducció*, (12):455–463.

Macketanz, V., Avramidis, E., Burchardt, A., Helcl, J., and Srivastava, A. (2017). Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. *Cybernetics and Information Technologies*, 17(2):28–43.

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

Popović, M. (2017). Comparing language related issues for nmt and pbmt between German and English. *The Prague Bulletin of Mathematical Linguistics*, 108(1):209–220.

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.

Specia, L., Harris, K., Blain, F., Burchardt, A., Macketanz, V., Skadiņa, I., Negri, M., and Turchi, M. (2017). Translation quality and productivity: A study on rich morphology languages. In *Proceedings of MT Summit XVI, the 16th Machine Translation Summit*, pages 55–71, Nagoya, Japan.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, pages 2214–2218, Istanbul, Turkey.

Toral, A. and Way, A. (2018). What level of quality can neural machine translation attain on literary text? In *Translation Quality Assessment*, pages 263–287. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Vaswani, A., Zhao, Y., Fossum, V., and Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1387–1392, Seattle, Washington, USA.

Yeh, A. (2000). More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING 2000*, pages 947–953, Saarbrücken, Germany.

A Supplementary Material

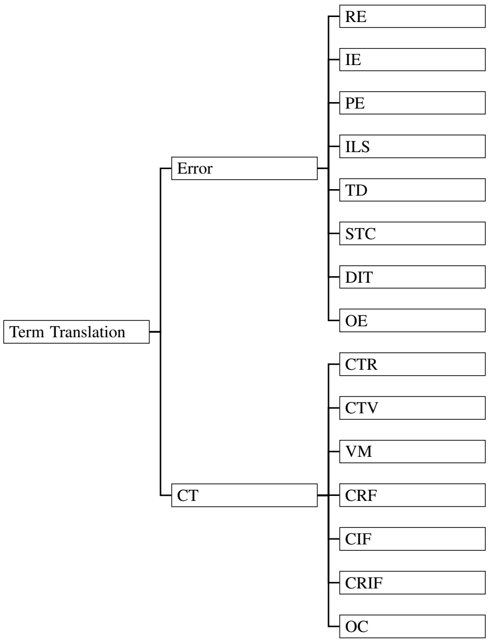
Table 5: PB-SMT vs NMT: English-to-Hindi.

	PB-SMT	NMT	∩	p-value
CTR	1,907	2,015	1662	
CTV	695	662	466	
VM	35	36	10	
CRF	4	7	4	
CIF	112	87	31	
CRIF				
OC	8	4		
CT	**2,761**	**2,811**	**2614**	
RE	15	5		0.044
IE	79	77	30	0.91
PE	52	47	19	0.61
ILS	77	44	9	0.001
TD	53	56	9	0.83
STC	12			
OE	12	24	2	
DIT	3			
ERROR	**303**	**253**	**86**	**0.011**

Table 6: PB-SMT vs NMT: Hindi-to-English.

	PB-SMT	NMT	∩	p-value
CTR	2,313	2,295	2,075	
CTV	241	245	147	
VM	24	33	5	
CRF	13	11	4	
CIF	75	107	48	
CRIF		2		
OC	2	18		
CT	**2,668**	**2,711**	**2,483**	
RE	18	5	1	0.008
IE	118	76	21	0.0009
PE	65	73	31	0.42
ILS	139	90	35	0.0001
TD	38	86	6	0.0001
STC	17			
OE		23		
DIT	1			
ERROR	**396**	**353**	**115**	**0.04**

Figure 1: Terminology Translation Typology.


RE
IE
PE
ILS
Error
TD
STC
DIT
Term Translation
OE
CTR
CTV
VM
CT
CRF
CIF
CRIF
OC