TermEval: An Automatic Metric for Evaluating Terminology Translation in MT

Rejwanul Haque, Mohammed Hasanuzzaman, and Andy Way

ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
{firstname.lastname}@adaptcentre.ie

Abstract. Terminology translation plays a crucial role in domain-specific machine translation (MT). Preservation of domain-knowledge from source to target is arguably the most concerning factor for the customers in translation industry, especially for critical domains such as medical, transportation, military, legal and aerospace. However, evaluation of terminology translation, despite its huge importance in the translation industry, has been a less examined area in MT research. Term translation quality in MT is usually measured with domain experts, either in academia or industry. To the best of our knowledge, as of yet there is no publicly available solution to automatically evaluate terminology translation in MT. In particular, manual intervention is often needed to evaluate terminology translation in MT, which, by nature, is a time-consuming and highly expensive task. In fact, this is unimaginable in an industrial setting where customised MT systems are often needed to be updated for many reasons (e.g. availability of new training data or leading MT techniques). Hence, there is a genuine need to have a faster and less expensive solution to this problem, which could aid the end-users to instantly identify term translation problems in MT. In this study, we propose an automatic evaluation metric, *TermEval*, for evaluating terminology translation in MT. To the best of our knowledge, there is no gold-standard dataset available for measuring terminology translation quality in MT. In the absence of gold standard evaluation test set, we semi-automatically create a gold-standard dataset from English–Hindi judicial domain parallel corpus.
We trained state-of-the-art phrase-based SMT (PB-SMT) and neural MT (NMT) models on two translation directions: English-to-Hindi and Hindi-to-English, and use TermEval to evaluate their performance on terminology translation over the created gold standard test set. In order to measure the correlation between TermEval scores and human judgments, translations of each source terms (of the gold standard test set) is validated with human evaluator. High correlation between TermEval and human judgements manifests the effectiveness of the proposed terminology translation evaluation metric. We also carry out comprehensive manual evaluation on terminology translation and present our observations.

Keywords: Terminology Translation · Machine Translation · Phrase-Based Statistical Machine Translation · Neural Machine Translation · Term Translation Evaluation

1 Introduction

Terms are productive in nature, and new terms are being created all the time. A term could have multiple meanings depending on the context in which it appears. For example,

words “terminal” (‘a bus terminal’ or ‘terminal disease’ or ‘computer terminal’) and “play” (‘play music’ or ‘plug and play’ or ‘play football’ or ‘a play’) could have very different meanings depending on the context in which they appear. A polysemous term (e.g. terminal) could have many translation equivalents in a target language. For example, the English word ‘charge’ has more than twenty target equivalents in Hindi (e.g. ‘dam’ for ‘value’, ‘bhar’ for ‘load’, ‘bojh’ for ‘burden’). While encountering a judicial document, translation of “charge” has to be a particular Hindi word: ‘aarop’. The target translation could lose its meaning if one does not take the term translation and domain knowledge into account. So the preservation of domain knowledge from source to target is pivotal in any translation workflow (TW), and this is one of the customers’ primary concerns in the translation industry. Naturally, translation service providers (TSPs) who use MT in production expect translations to be consistent with the relevant context and the domain in question. However, evaluation of terminology translation has been one of the least explored arena in MT research. No standard automatic MT evaluation metric (e.g. BLEU [35]) can provide much information on how good or bad a MT system is in translating domain-specific expressions. To the best of our knowledge, as of now no one has proposed any effective way to automatically evaluate terminology translation in MT. In industrial TWs, in general, TSPs hire manual experts relating to the concerned domain for identifying term translation problems in MT. Nevertheless, such human evaluation process is normally laborious, expensive and time-consuming task. Moreover, in an industrial setting, retraining of customer-specific MT engine from scratch is carried out quite often when a reasonable size of new training data pertaining to the domain and styles of that on which that MT system was built or a new state-of-the-art MT technique is available. In industry, carrying out human evaluation on term translation from scratch each time when a MT system is updated is an unimaginable task in a commercial context. This is an acute problem in industrial TW and renowned TSPs want it to be solved for their own interest. A suitable solution to the problem of terminology translation evaluation would certainly aid the MT users who want to quickly assess theirs MT systems in the matter of the domain-specific term translation, and be a blessing for the translation industry.

In this work, we propose an automatic evaluation metric, TermEval, to quickly assess terminology translation quality in automatic translation. With this, we aim to provide the MT users a solution to the problem of the terminology translation evaluation. The proposed automatic evaluation metric TermEval, we believe, would ease the problem of those MT users who often need to carry out evaluation on terminology translation. Since there is no publicly available gold standard for term translation evaluation, we create a gold-standard evaluation test set from a legal domain data (i.e. judicial proceedings) following a semi-automatic terminology annotation strategy. We use our inhouse bilingual term annotation tool, *TermMarker* (cf. Section 4). In short, TermMarker marks source and target terms in either side of a test set, incorporate lexical and inflectional variations of the terms relevant to the context in which they appear, with exploiting the automatic terminology extraction technique of [17, 18]. The annotation technique needs little manual intervention to validate the term tagging and mapping, provided a rather noisy automatic bilingual terminology in annotation interface. In an industrial set-up, TSPs would view this method as an ideal and onetime solution to the one that we pointed

out above since the annotation scheme is to be a cheaper and faster exercise and will result a reusable gold standard in measuring term translation quality.

PB-SMT [30], a predictive modeling approach to MT, was the main paradigm in MT research for more than two decades. NMT [25, 9, 47, 4, 50], an emerging prospect for MT research, is an approach to automatic translation in which a large neural network (NN) is trained by deep learning techniques. Over the last six years, there has been incremental progress in the field of NMT to the point where some researchers are claiming to have parity with human translation [19]. Nowadays, NMT, despite being a relatively new addition to this field of research, is regarded as a preferred alternative to previous mainstream methods and represents a new state-of-the-art in MT research. We develop competitive PB-SMT and NMT systems with a less examined and low-resource language pair, English–Hindi. Hindi is a morphologically rich and highly inflected Indian language. Our first investigation is from a less inflected language to a highly inflected language (i.e. English-to-Hindi), and the second one is the other way round (i.e. Hindi-to-English). With this, we compare term translation in PB-SMT and NMT with a difficult translation pair involving two morphologically divergent languages. We use TermEval to evaluate MT system's performance on terminology translation on gold standard test set. To check how TermEval correlates with the human judgements, the translation of each source term (of the gold standard test set) is validated with human evaluators. We found that TermEval represents a promising metric for automatic evaluation of terminology translation in MT, with showing very high correlation with the human judgements. In addition, we demonstrate a comparative study on term translation in PB-SMT and NMT in two set-ups: automatic and manual. To summarize, our main contributions in this paper are as follows:

1. We semi-automatically create a gold standard evaluation test set for evaluating terminology translation in MT. We demonstrate various linguistic issues and challenges in relation to the annotation scheme.
2. To the best of our knowledge, we are the first to propose an automatic evaluation metric for evaluating terminology translation in MT, namely TermEval. We test TermEval and found it be an excellent in quality while measuring term translation errors in MT.
3. We compare PB-SMT and NMT in terminology translation on two translation directions: English-to-Hindi and Hindi-to-English. We found that NMT is less error-prone than PB-SMT with regard to the domain-specific terminology translation.

The remainder of the paper is organised as follows. In Section 2, we discuss related work. Section 3 describes our MT systems used in our experiments. In Section 4, we present how we created gold standard dataset and examine challenges in relation to the termbank creation process. In Section 5, we report our evaluation plan and experimental results. Section 6 describes discussion and analysis while Section 7 concludes, and provides avenues for further work.

2 Related Work

2.1 Terminology Annotation

Annotation techniques have been widely studied in many areas of natural language processing (NLP). To the best of our knowledge, no one has explored the area in relation to the term annotation in corpus barring [37] who investigated term extraction, tagging and mapping techniques for under-resourced languages. They mainly present methods for term extraction, term tagging in documents, and bilingual term mapping from comparable corpora for four under-resourced languages: Croatian, Latvian, Lithuanian, and Romanian. The paper primary focused on acquiring bilingual terms from comparable Web crawled narrow domain corpora similar to the study of [17, 18] who automatically create bilingual termbank from parallel corpus.

In our work, we select a test set from a (judicial) domain parallel corpus, and semi-automatically annotate the test set sentences by marking source and target terms in either side of that and incorporating lexical and inflectional variations of the terms relevant to the context in which they appear. For annotation we took support from a rather noisy bilingual terminology that was automatically created from juridical corpus. For automatic bilingual term extraction we followed the approach of [17, 18].

2.2 Term Translation Evaluation Method

As far as measuring terminology translation quality in MT is concerned, researchers or end-users (e.g. translation industry) generally carry out manual evaluation with domain experts. Farajian et al. [13] proposed an automatic terminology translation evaluation metric which computes the proportion of terms in the reference set that are correctly translated by the MT system. This metric looks for source and target terms in the reference set and translated documents, given a termbank. There is a potential problem with this evaluation method. There could be a possible instance where a source term from the input sentence is incorrectly translated into the target translation, and the reference translation of the source term spuriously appears in the translation of a different input sentence. In such cases, the above evaluation method would make hit counts which are likely to be incorrect. In addition to the above problem, there are two more issues that [13] cannot address, which we discuss in Section 4.4 where we highlight the problem with the translation of ambiguous terms and in Section 4.2 where we point out the consideration of the lexical and inflectional variations for a reference term, given the context in which the reference translation appears.

In order to evaluate the quality of the bilingual terms in MT, Arčan et al. [2] manually created a terminology gold standard for the IT domain. They hired annotators with a linguistic background to mark all domain-specific terms in the monolingual GNOME and KDE corpora [48]. Then, the annotators manually created a bilingual pair of two domain-specific terms found in a source and target sentence, one being the translation of the other. This process resulted in the identification of 874 domain-specific bilingual terms in the two datasets [1]. The end goal (i.e., evaluating the quality of term translation in MT) of their manual annotation task was identical to that of this study. However, our annotation task is a semi-automatic process that helps create a terminology gold standard

more quickly. In this work, we intent to ease this problem with proposing an automatic term translation evaluation metric. In short, the annotation task takes support from a bilingual terminology that is automatically created from a bilingual domain corpus. For automatic bilingual term extraction, we followed the approach of [17, 18]. In this context, an obvious challenge in relation to the term annotation task is that there is a lack of a clear definition of terms (i.e., what entities can be labelled as terms [36]). While it is beyond the scope of this article to discuss this matter, the various challenges relating to terminology annotation, translation and evaluation will be presented in more detail.

2.3 PB-SMT versus NMT: Terminology Translation & Evaluation

Since the introduction of NMT to MT community, researchers investigate to what extents and in what aspects NMT are better (or worse) than PB-SMT. In a quantitative comparative evaluation, [24] report performance of PB-SMT and NMT across fifteen language pairs and thirty translation directions on the United Nations Parallel Corpus v1.0 [55] and show that for all translation directions, NMT is either on par with or surpasses PB-SMT. We refer interested reader [49] for more works in this direction. In a nutshell, most of the studies in this direction show that the NMT systems provide better translation quality (e.g. more fluent, less lexical, reordering and morphological errors) than the PB-SMT systems.

We now turn our particular attention to the papers that looked into terminology translation in NMT. Burchardt et al. [7] conduct a linguistically driven fine-grained evaluation to compare rule-based, phrase-based and neural MT engines for English–German based on a test-suite for MT quality, confirming the findings of previous studies. In their German-to-English translation task, PB-SMT, despite reaching the lowest average score, is the best-performing system on named-entities and terminology translation. However, when tested on reverse translation direction (i.e. English-to-German), a commercial NMT engine becomes the winner as long as the term translation is concerned. In a similar experimental set-up, Macketanz et al. [33] report that their PB-SMT system outperforms NMT system on terminology translation on both in-domain (IT domain) and general domain test suites in an English-to-German translation task. Specia et al. [45] carried out an error annotation process using the Multidimensional Quality Metrics error annotation framework [32] on MT PE environment. The list of errors is divided into three main categories: accuracy, fluency and terminology. According to the annotation results, terminology-related errors are found more in NMT translations than in PB-SMT translations in English-to-German task (139 vs 82), and other way round in English-to-Latvian task (31 vs 34). Beyer et al. [5], from their manual evaluation procedure, report that PB-SMT outperforms NMT on term translation, which they speculate could be because their technical termbank was part of the training data used for building their PB-SMT system. Špela [52] conducts an automatic and small-scale human evaluation on terminology translation quality of Google Translator NMT model [53] compared to its earlier PB-SMT model for Slovene–English language pair and in the specialised domain of karstology. The evaluation result of Slovene-to-English task confirms NMT is slightly better than PB-SMT in terminology translation, while the opposite direction (i.e. English-to-Slovene task) shows a reversed picture with PB-SMT outperforming NMT. Špela [52] carries out a little bit qualitative analysis, with counting

terms that are dropped in target translations and detailing instances where MT systems often failed to preserve the domain knowledge. More recently, we investigated domain term translation in PB-SMT and NMT on English<=>Hindi translation tasks [16] and found that the NMT systems commit fewer lexical, reordering and morphological errors than the PB-SMT systems. The opposite picture is observed in the case of term omission in translation, with NMT omitting more terms in translation than PB-SMT.

As far as terminology translation quality evaluation in PB-SMT alone is concerned, given its relatively a longer history, there has been many papers that has investigated this problem. For an example, [22] investigated term translation in a PB-SMT task and observed that more than 10% of high-frequency terms are incorrectly translated by their PB-SMT decoder, although the system's BLEU [35] score is quite high, i.e. 63.0 BLEU. One common thing in the papers [22, 7, 33, 45, 5, 52], above is that they generally carried out subjective evaluation in order to measure terminology translation quality in MT, which, as mentioned earlier, is a time-consuming and expensive task. In this paper, we propose an automatic evaluation strategy that automatically determines terminology translation errors in MT. We observe terminology translation quality in PB-SMT and NMT on two translation directions: English-to-Hindi and Hindi-to-English, and two set-ups: automatic and subjective evaluations.

3 MT Systems

3.1 PB-SMT system

For building our PB-SMT systems we used the Moses toolkit [29]. We used 5-gram LM trained with modified Kneser-Ney smoothing using KenLM toolkit [21]. 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 NPLM toolkit[1] [51] on the target-side of parallel training corpus alone. Our PB-SMT log-linear features include: (a) 4 translational features (forward and backward phrase and lexical probabilities), (b) 8 lexicalised reordering probabilities (*wbe-mslr-bidirectional-fe-allff*), (c) 2 5-gram LM probabilities (Kneser-Ney and NPLM), (d) 5 OSM features [12], and (e) word-count and distortion penalties. In our experiments word alignment models are trained using GIZA++ toolkit[2] [34], phrases are extracted following *grow-diag-final-and* algorithm of [30], Kneser-Ney smoothing is applied at phrase scoring, and a smoothing constant (0.8u) is used for training lexicalized reordering models. The weights of the parameters are optimized using the margin infused relaxed algorithm [8] on the development set. For decoding the cube-pruning algorithm [23] is applied, with a distortion limit of 12. We call the English-to-Hindi and Hindi-to-English PB-SMT systems EHPS and HEPS, respectively.

3.2 NMT system

For building our NMT systems we used the MarianNMT [24] toolkit. The NMT systems are Google transformer models [50]. In our experiments we followed the recommended

[1] https://www.isi.edu/natural-language/software/nplm/
[2] http://www.statmt.org/moses/giza/GIZA++.html

best set-up from [50]. The tokens of the training, evaluation and validation sets are segmented into sub-word units using the Byte-Pair Encoding (BPE) technique [14], which was proposed by [42]. Since English and Hindi are written in Roman and Devanagari scripts and have no overlapping characters, the BPE is applied individually on the source and target languages. We performed 32,000 join operations. Our training set-up is detailed below. We consider size of encoder and decoder layers 6. As in [50], we employ residual connection around layers [20], followed by layer normalisation [3]. The target embeddings and output embeddings are tied in output layer [40]. Dropout [15] between layers is set to 0.10 . We use mini-batches of size 64 for update. The models are trained with Adam optimizer [26], with learning-rate set to 0.0003 and reshuffling the training corpora for each epoch. As in [50], we also use the learning rate warm-up strategy for Adam. The validation on development set is performed using three cost functions: cross-entropy, perplexity and BLEU. The early stopping criteria is based on cross-entropy, however, the final NMT system is selected as per highest BLEU scores on the validation set. The beam size for search is set to 12.

Initially, we use parallel training corpus to build our English-to-Hindi and Hindi-to-English baseline transformer models. We translate monolingual sentences (cf. Table 1) with the baseline models and create source synthetic sentences [41]. Then, we append this synthetic training data to the parallel training data, and retrain the baseline models. We make our final NMT model with ensembles of 4 models that are sampled from the training run. We call our final English-to-Hindi and Hindi-to-English NMT systems EHNS and HENS, respectively.

3.3 Data Used

For experimentation we used the IIT Bombay English-Hindi parallel corpus[3] [31] that is compiled from a variety of existing sources, e.g. OPUS[4] [48]. That is why the parallel corpus is a mixture of various domains. For building additional language models (LMs) for Hindi and English we use the HindEnCorp monolingual corpus [6] and monolingual corpus from various sources (e.g. the European Parliamentary proceedings [28]) from the OPUS project, 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. In order to perform tokenisation for English and Hindi, we used the standard tokenisation tool[5] of the Moses toolkit.

3.4 PB-SMT versus NMT

In this section, we present the comparative performance of the PB-SMT and NMT systems in terms of automatic evaluation metrics: BLEU, METEOR [11], TER [44],

[3] http://www.cfilt.iitb.ac.in/iitb_parallel/

[4] http://opus.lingfil.uu.se/

[5] https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl

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

Table 1. Corpus Statistics. English (EN), Hindi (HI).

chrF3 [38] and BEER [46]. The BLEU, METEOR and TER are standard metrics which are widely used by the MT community. The chrF3 and BEER metrics, which are character-level n-gram precision-based measures, are proved to have high correlation with human evaluation. Note that TER is an error metric, which means lower values indicate better translation quality. We report evaluation results in Table 2.

	BLEU	METEOR	TER	chrF3	BEER
EHPS	28.8	30.2	53.4	0.5247	0.5972
EHNS	36.6	33.5	46.3	0.5854	0.6326
HEPS	34.1	36.6	50.0	0.5910	0.6401
HENS	39.9	38.5	42.0	0.6226	0.6644

Table 2. Performance of PB-SMT and NMT systems on automatic evaluation metrics.

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 a difficult translation pair. These BLEU scores, in fact, underestimate the translation quality, given the relatively free word order in Hindi, providing just single reference translation set for evaluation. Many TSPs consider 30.0 BLEU score as a benchmarking value and use those MT systems in their TW that produce BLEU scores above the benchmarking value. For example, [43] successfully used an English-to-Latvian MT system with a similar BLEU score (35.0 BLEU) in SDL Trados CAT tool.[6] In this perspective, EHPS is just below par and EHNS is well above the benchmarking value.

[6] https://en.wikipedia.org/wiki/SDL_Trados_Studio

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, MT quality from the morphologically-rich to morphologically-poor language improves. The differences in BLEU scores of PB-SMT and NMT systems in both the English-to-Hindi and Hindi-to-English translation tasks are statistically significant [27]. This trend is observed with the remaining evaluation metrics.

4 Creating Gold Standard Evaluation Set

This section describes a technique that semi-automatically creates a gold standard evaluation test data for evaluating terminology translation in MT. Our proposed automatic evaluation metric, TermEval, is based on the gold standard test set. To exemplify, we present the semi-automatic test data creation technique for the English–Hindi language pair below. For evaluating term translation with our MT systems (cf. Section 3) we use the test set (cf. Table 1) that contains 2,000 sentence-pairs and is from judicial domain. We annotated the test set by marking term-pairs on the source- and target-sides of the test set. The annotation process is accomplished with our own bilingual term annotation tool, TermMarker. It is an user-friendly GUI developed with PyQT5.[7] We have made TermMarker publicly available to the research community via a software repository.[8] The annotation process starts with displaying a source–target sentence-pair from the test set at TermMarker's interface. If there is a source term present in the source sentence, its translation equivalent (i.e. target term) is found in the target sentence, then the source–target term-pair is marked. The annotation process is simple, which is carried out manually. The annotator, who is a native Hindi evaluator with excellent English skills, is instructed to mark those words as terms that belong to legal or judicial domains. The annotator was also instructed to mark those sentence-pairs from the test set that contains errors (e.g. mistranslations, spelling mistakes) in either source or target sentences. The annotator reported 75 erroneous sentence-pairs which we discard from the test set. In addition to this, there are 655 sentence-pairs of the test set that do not contain any terms. We call the set of remaining 1,270 sentence-pairs *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.[9]

4.1 Annotation Suggestions from Bilingual Terminology

TermMarker supports annotation suggestions from an external terminology, if supplied. We recommend this option for faster annotation, although this is optional. For example, in our case, while manually annotating bilingual terms in the judicial domain test set we took support from a rather noisy bilingual terminology that was automatically created from the Judicial corpus (cf. Table 1). For automatic bilingual term extraction we followed the benchmark approach of [17, 18] which is regarded as the state-of-the-art terminology extraction technique and works well even on as few as 5,000 parallel

[7] https://en.wikipedia.org/wiki/PyQt

[8] `https://github.com/rejwanul-adapt/TermMarker`

[9] `https://github.com/rejwanul-adapt/EnHiTerminologyData`

segments. The user chooses one of the three options for an annotation suggestion: accept, skip and reject. The rejected suggestion is excluded from the bilingual terminology to make sure it never appears as the annotation suggestion in future. The newly marked term-pair is included in the bilingual termbank, which are to be used in annotation process.

Number of Source–Target Term-pairs		3,064
English	Terms with LIVs	2,057
	LIVs/Term	5.2
Hindi	Terms with LIVs	2,709
	LIVs/Term	8.4

Table 3. Statistics of occurrences of terms in gold-testset.

In Table 3, we show statistics of occurrences of terms in the gold standard evaluation set (i.e. gold-testset). We found 3,064 English terms and their target equivalents (3,064 Hindi terms) in source- and target-sides of gold-testset, respectively, i.e. the number of aligned English–Hindi term-pairs in gold-testset is 3,064. We observed presence of the nested terms (i.e. overlapping terms) in gold-testset, e.g. 'oral testimony' and 'testimony', 'pending litigation' and 'litigation', 'attesting witness' and 'witness'. In nested terms, we call a higher-gram overlapping term (e.g. 'oral testimony') a *superterm*, and a lower-gram overlapping term (e.g. 'testimony') a *subterm*. A nested term may have more than one subterm, but it has to have only one superterm. TermMarker allows us to annotate both subterms and superterms.

4.2 Variations of Term

A term could 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 English word 'affidavit' has multiple target equivalents (lexical and inflectional variations) in Hindi even if the translation domain is legal or juridical: 'shapath patr', 'halaphanaama', 'halaphanaame', '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). In similar contexts, translation of the English preposition 'of' has multiple variations (postpositions) ('ka', 'ke') in Hindi. For this, an English term 'thumb impression' may have many translations in Hindi, e.g. 'angoothe **ka** nishaan' and 'angoothe **ke** nishaan', where 'angoothe' means 'thumb' and 'nishaan' means 'impression'.

For each term we check whether the term has any additional variations (lexical or inflectional) pertaining to the juridical domain and relevant to the context of the sentence. If this is the case, we include the relevant variations as the alternatives of the term. The idea is to create termbank as exhaustive as possible. In Table 3, we report the number of

English and Hindi terms for which we added lexical and inflectional variations (LIVs), and the average number of variations per such term. As expected, both the numbers are higher in Hindi than English.

During annotation, the user can manually add relevant variations for a term through TermMarker's interface. However, we again exploit the method of [17, 18] for obtaining variation suggestions for a term. The automatically extracted bilingual terminology of [17, 18] comes with the four highest-weighted target terms for a source term. If the user accepts an annotation suggestion (source–target term-pair) from the bilingual terminology, the remaining three target terms are considered as the variation suggestions of the target term. Like the case of an annotation suggestion above, the user chooses one of the three options for a variation suggestion: accept, skip and reject. The rejected variation is excluded from the bilingual terminology to make sure it never appears as a variation suggestion in future. The newly added variation is included in the bilingual terminology for future use. Note that TermMarker has also an option to conduct annotation in both ways (source-to-target and target-to-source) at the same time. For this, the user can optionally include a target-to-source bilingual terminology.

4.3 Consistency in Annotation

As pointed out in the sections above, new term-pairs and variations of terms are often added to the terminology at the time of annotation. This may cause inconsistency in annotation since new term-pairs or variations could be omitted for annotation in the preceding sentences that have already been annotated. In order to eliminate the inconsistency problem, TermMarker includes a *checkup module* that traces annotation history, mainly with storing rejected and skipped items. The checkup module notifies the human annotator when any of the preceding sentences has to be annotated for a newly included term-pair or variation of a term to the termbank.

On completion of the annotation process, a set of randomly selected 100 sentence-pairs from gold-testset were further annotated for agreement analysis. Inter-annotator agreement was computed using Cohen's kappa [10] at word-level. This means for a multi-word term we consider number of words in it for this calculation. 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 to be excellent in quality.

4.4 Ambiguity in Terminology Translation

One can argue that term annotation can be accomplished automatically if a bilingual terminology is available for the target domain. If we automatically annotate a test set with a given terminology, the automatic annotation process will likely to introduce noise into the test set. As an example, the translation of an English legal term 'case' is 'mamla' in Hindi. The translation of the word 'case' could be the same Hindi word (i.e. 'mamla') even if the context is not legal. A legal term 'right' can appear in a legal/juridical text with a completely different context (e.g. fracture of right elbow). The automatic annotation process will ignore the contexts in which these words ('case', 'right', 'charge') belong

and incorrectly mark these ambiguous words as terms. The automatic annotation process will introduce even more noise while adding variations for a term from a termbank or other similar sources (e.g. dictionary) for the same reason.

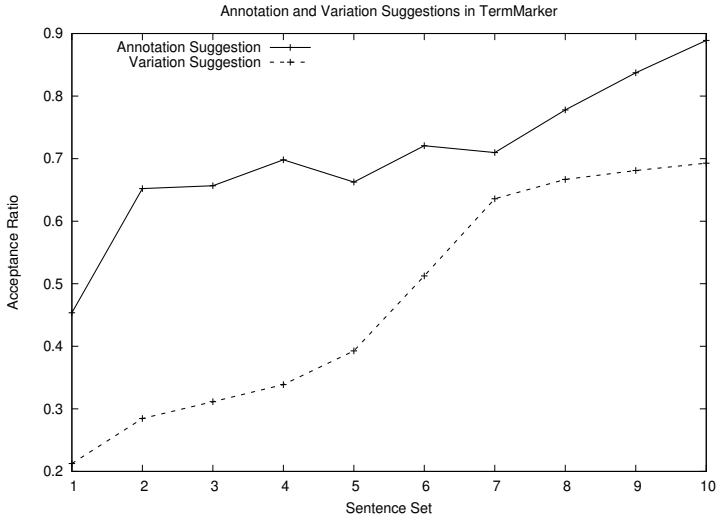

Fig. 1. Curves for Acceptance Ratio of Suggestions.

4.5 Measuring Performance of TermMarker

We tested whether the use of annotation and variation suggestions from bilingual terminology [17, 18] makes the annotation process more faster. For this, we divided our judicial domain test set (2,000 segments, cf. Table 1) into 10 sets, each of which contains 200 segment-pairs. Note that we started our annotation process on set 1 and ended on set 10. We count the number of total annotation and variation suggestions as well as the number of accepted annotation and variation suggestions over each set. We plot the ratio of these (e.g. accepted annotation suggestions / total annotation suggestions) against the segment set number in Figure 1. The x-axis and y-axis of Figure 1 represent the segment set number and acceptance ratio, respectively. We see from Figure 1 that both curves (annotation and variation suggestions) move upward over time, i.e. the acceptance rate of suggestions increases over time. This is because anomalous entries are rejected from the terminology and new valid entries are added into the terminology all the time, which makes the annotation process iteratively faster.

5 Evaluating Terminology Translation in MT

5.1 Automatic Evaluation: TermEval

The proposed evaluation metric starts evaluation process with forming a tuple with each source sentence from the test set, its translation (i.e. hypothesis), the list of source terms appearing in the source sentence, their reference translations (i.e. terms), and LIVs of the reference terms. First, we look for the reference terms (or the LIVs of the reference terms) in the hypothesis.[10] If a reference term (or one of its LIVs) is found in the hypothesis, this indicates that the MT system has correctly translated the source term into the target language. If this is not the case, there is likely to be an error in that term translation. At the end of the iteration (over all tuples), we obtain the TermEval score. The evaluation method includes an additional procedure that takes nested overlapping terms and multiple identical reference terms (or variations) into account. More formally, TermEval is calculated provided the test and translation sets using (1):

$$\text{TermEval} = \frac{\sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{v=1}^{V} \begin{cases} 1 & \text{if } \mathrm{R}_v \in \mathrm{Hyp}_n;\ break; \\ 0 & \text{otherwise} \end{cases}}{\text{NT}} \tag{1}$$

where: N : number of sentences in the test set
S : number of source terms in the n^{th} source sentence
V : number of reference translations (including LIVs) for s^{th} the source term
R_v : v^{th} reference term for s^{th} the source term
Hyp_n : translation of n^{th} source input sentence
NT : total number of terms in the test set

5.2 Term Translation Accuracy with TermEval

In this section we obtain evaluation results to evaluate the performance of the MT models on terminology translation using gold-testset that includes 3,064 source–target term-pairs. The terminology translation quality is measured using the TermEval metric (cf. (1)), and the scores are reported in Table 4. For clarity, we also report the total number of correct translation (CT) in the table, which, in fact, is the numerator of the right side of (1). As can be seen from the table, EHPS and EHNS correctly translate 2,610 and 2,680 English terms (out of total 3,064 terms), respectively, into Hindi, resulting the TermEval

[10] Since Moses can supply word-to-word alignments with its output (i.e. translation) from the phrase table (if any), one can exploit this information to trace target translation of a source term in the output. However, there are few potential problems with the alignment information, e.g. there could be null or erroneous alignments. Note that, at the time of this work, the transformer models of MarianNMT could not supply word-alignments (i.e. attention weights). In fact, our intention is to make our proposed evaluation method as generic as possible so that it can be applied to the output of any MT system (e.g. an online commercial MT engine). This led us to abandon such dependency.

	CT	TermEval
EHPS	2,610	0.852
EHNS	2,680	0.875
HEPS	2,554	0.834
HENS	2,540	0.829

Table 4. PB-SMT vs NMT: Terminology Translation Accuracy in TermEval.

scores of 0.852 and 0.875, respectively. We use approximate randomization [54] to test the statistical significance of the difference between two systems. We found that the difference between the scores is statistically significant. On the other direction (i.e. from Hindi-to-English), HEPS and HENS correctly translate 2,554 and 2,540 English terms (out of total 3,064 terms), respectively, into Hindi, resulting TermEval of 0.834 and 0.829, respectively. Unlike the above, we found that the difference between the scores are not statistically significant.

	LIVs	%
EHPS	703	26.0
EHNS	665	24.5
HEPS	241	11.7
HENS	245	11.9

Table 5. Number of seen LIVs in translations.

Proportion of LIVs Seen in Translation We counted the number of instances where a source term is correctly translated into target translation and the translation-equivalent of that term is one of the variations of the reference term; the percentage with respect to the total number of reference terms which includes variations (cf. Table 3) in the gold-testset is shown in Table 5. We see from the table that these numbers are much higher in the English-to-Hindi PB-SMT task (26.0% and 24.5%) compared to the Hindi-to-English task (11.7% and 11.9%). Hindi is a morphologically rich and highly inflected language, and we see from Table 1 that the training set vocabulary size is much higher in Hindi compared to English. This could be the reason why the numbers are much higher in the English-to-Hindi task.

5.3 Manual Evaluation Method

This section presents our manual evaluation plan. As mentioned above, the sentences of gold-testset were translated with the English-to-Hindi and Hindi-to-English MT systems (cf. Section 3). Translations of the source terms of gold-testset were manually validated and classified into two categories: error and correct. This was accomplished

with the human evaluators. 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 source and reference sentences, respectively, and displays the LIVs of the reference term, if any. The GUI lists the correct and incorrect categories. The evaluator, a native Hindi speaker with the excellent English and Hindi skills, is instructed to follow the following criterion 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 on GUI, (b) do not need to judge whole translation instead look at the local context in which both source term and its translation belong to, 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 randomly selected additional 100 source terms for further classification. The idea is to measure agreement in manual classification. We considered the binary categories (correct or incorrect term translation) in calculation, i.e. we count an agreement whenever both evaluators agree that it is correct (or incorrect) term translation, with agreement by chance = 1/2. As far as the agreements on classification of terminology translations with the four MT systems are concerned, we found that the kappa coefficient for the binary classes ranges from 0.97 to 1.0. It is believed that a kappa coefficient between 0.6 – 0.8 represents substantial agreement, with anything above 0.8 indicating perfect agreement. In this sense, our manual term translation classification quality can be labeled as excellent.

5.4 Measuring Term Translation Accuracy from Manual Classification Results

	CT	ACC
EHPS	2,761	0.901
EHNS	2,811	0.917
HEPS	2,668	0.870
HENS	2,711	0.884

Table 6. PB-SMT vs NMT: Terminology Translation Accuracy (Manual). CT: number of correct translation, ACC: Accuracy.

Given the manual classification results above, we obtain the performance of our MT systems on terminology translation. We measure term translation accuracy given the number of correct term translations and the total number of terms in gold-testset, i.e. 3,064. As in above, we report the total number of correct translations (CT) in Table 6. For comparison, we also report accuracy (ACC) (i.e. this is measured as the fraction of the correct term translations over the total number of term translations) in Table 6. As can be seen from the table, EHPS and EHNS correctly translate 2,761 and 2,811 English terms (out of total 3,064 terms), respectively, into Hindi, resulting ACC of 0.901 and

0.917, respectively. We use approximate randomization to test the statistical significance of the difference between two systems. We found that the difference between the scores is statistically significant. In the case of the Hindi-to-English task, we see that HEPS and HENS correctly translate 2,668 and 2,711 English terms (out of total 3,064 terms), respectively, into Hindi, resulting ACC of 0.870 and 0.884, respectively. As in above, the difference between the scores is statistically significant.

5.5 Overlapping Correct and Incorrect Term Translations

	PB-SMT	NMT	PB-SMT ∩ NMT
	English-to-Hindi Task		
Correct	2,761	2,811	2614
Incorrect	303	253	86
	Hindi-to-English Task		
Correct	2,668	2,711	2,483
Incorrect	396	353	115

Table 7. Number of Overlaps in Correct and Incorrect Translation in PB-SMT and NMT

Toral and Sánchez-Cartagena [49] carried out a comparative analysis with measuring overlap between an output by NMT and another by PB-SMT systems. Likewise, we report the number of pairwise overlap, i.e. number of instances in which NMT and PB-SMT have identical manual term classification outcomes (correct). The last columns of Table 7 show the numbers of pairwise overlap. The number of overlapping instances under the incorrect type are 86 and 115 that are nearly one third of the total numbers of error committed by the PB-SMT and NMT systems alone, indicating majority of the errors in PB-SMT are complementary with those in NMT. This finding on terminology translation is corroborated with that of [39] who finds complementarity with the issues relating to the translations of NMT and PB-SMT.

5.6 Validating TermEval

This section presents how we measure the correlation of human evaluation and automatic evaluation in the matter of terminology translation in MT.

Contingency Table Using the TermEval metric we obtained the number of correct (and incorrect) term translations by the MT systems, which were reported in Table 4. With the human evaluation we obtain the actual number of correct (and incorrect) term translations by the MT systems on the test set, which were shown in Table 6. Therefore, given the automatic and human evaluation results, it is straightforward for us to evaluate our proposed evaluation metric, TermEval. Hence, given the automatic and manual evaluation results, we create contingency tables for both the English-to-Hindi

English-to-Hindi Task

PB-SMT			NMT		
	2,610	454		2,680	384
2,761	2,602	159	2,811	2,677	134
303	8	295	253	3	250

Hindi-to-English Task

PB-SMT			NMT		
	2,554	510		2,540	524
2,668	2,554	114	2,711	2,540	171
396	0	396	353	0	353

Table 8. Contingency Tables.

and Hindi-to-English tasks and show them is Table 8. As can be seen from Table 8, there are two contingency tables for each task, left side tables are for the PB-SMT tasks and the right side tables are for the NMT tasks. The first row and column of each table represent automatic and manual classes, respectively. Each row or column shows two numbers, denoting the correct and incorrect term translations by the MT systems. Thus, the numbers from the manual classes are distributed over the automatic classes, and the other way round. For an example, manual evaluator labels 2,761 term translations (out of 3,064 total source terms) as correct translations in the English-to-Hindi PB-SMT task, and these 2,761 term translations belong to 2 categories (correct: 2,602, and incorrect: 159) as per the automatic evaluation.

Measuring Accuracy Give the contingency tables (cf. Table 8), we measure accuracy of TermEval in the English–Hindi translation task. For measuring accuracy we make use of three widely used metrics: precision, recall and F1. We report the scores in Table 9. As can be seen from Table 9, we obtain roughly similar scores in four translation tasks (i.e. ranging from F1 of 0.967 to F1 of 0.978), and generally very high precision and slightly low recall scores in all tasks. TermEval represents a promising metric as per the scores in Table 9. We provide our in-depth insights on the results from Table 9 below (cf. Section 6).

6 Discussion and Analysis

In this section, first we discuss the scenario in which TermEval labeled those term translations as correct, which, are, in fact, incorrect as per the human evaluation results (false positives, cf. Table 8). Then, we discuss the reverse scenario in which TermEval labeled those term translations as incorrect, which, are, in fact, correct as per the human evaluation results (false negatives, cf. Table 8).

6.1 False Positives

As can be seen from fifth row of Table 8, there are 8 and 3 false-positives in the English-to-Hindi PB-SMT and NMT tasks, respectively. This indicates that in each case

	EHPS	EHNS
P	0.997	0.999
R	0.942	0.953
F1	0.968	0.975
	EHPS	EHNS
P	1.0	1.0
R	0.957	0.937
F1	0.978	0.967

Table 9. Accuracy of Proposed Automatic Term translation Evaluation Method: precision, recall and F1 metrics.

TermEval labels the term translation as correct, because the corresponding reference term (or one of its LIVs) is found in hypothesis, although the manual evaluator labeled that term translation as incorrect. We verify these cases in translations with the corresponding reference terms and theirs LIVs. We found that in 8 cases out of 11 cases the LIV lists contain incorrect inflectional variations for the reference terms. In each of these cases, an inflectional variation of the reference term misfits in the context of the reference translation when used in the place of the corresponding reference term. These incidents can be viewed as annotation errors as these erroneous inflectional variations for the reference terms were included in gold-testset at the time of its creation. For the 3 remaining cases we found that the English-to-Hindi PB-SMT system made correct lexical choice for the source terms, although the meanings of theirs target-equivalents in the respective translation are different to those of the source terms. This can be viewed as a cross-lingual disambiguation problem. For an example, one of the three source terms 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'. It would be a challenging task for any evaluation metric as was the case for TermEval to correctly recognise such term translation errors. We keep this topic as a subject of future work.

6.2 False Negatives

We see from Table 8 that the number of false negatives (e.g. 159 in the English-to-Hindi PB-SMT task) across all MT tasks are much higher than that of false positives. This is, in fact, responsible for the slightly worse recall scores (cf. Table 9). We point out below why TermEval failed to label such term translations (e.g. 159 terms in the English-to-Hindi PB-SMT task) as correct despite the fact those are correct translations as per the human evaluation results.

Reordering Issue Here, first we highlight the word ordering issue in term translation. For an example, a Hindi source term "khand nyaay peeth ke nirnay" (English reference

term: “division bench judgment”) is correctly translated into the following English translation by the Hindi-to-English NMT system: “it shall also be relevant to refer to article 45 - 48 of the *judgment of the division bench*”. Nevertheless, TermEval implements a simple word matching module that, essentially, failed to capture such word ordering at target translation. In Table 10, we report the number of instances where TermEval failed to distinguish those term translation in the PB-SMT and NMT tasks that contains all words of the reference term (or one of its LIVs) but in a order different to the reference term (or one of its LIVs). As can be seen from Table 10, these numbers are slightly high when the target language is English. In order to automatically capture a term translation whose word-order is different to that of the reference term (or one of its LIVs), we need to incorporate language-specific or lexicalised reordering rules into TermEval, which we intend to investigate in future.

False Negative [Due to Term Reordering]			
	PB-SMT	NMT	PB-SMT ∩ NMT
English-to-Hindi	4	7	4
Hindi-to-English	13	11	4

Table 10. False negatives in the PB-SMT and NMT tasks when TermEval failed to distinguish reordered correct term translation.

Inflectional Issue We start the discussion with an example from the Hindi-to-English translation task. There is a source Hindi term ‘abhikathan’, its reference term is ‘allegation’, and a portion of the reference translation is ‘an allegation made by the respondent ...’. The LIV list of the reference term includes two lexical variations for ‘allegation’: ‘accusation’ and ‘complaint’. A portion of the translation produced by the Hindi-to-English NMT system is ‘it was *alleged* by the respondent ...’, where we see the Hindi term ‘abhikathan’ is translated into ‘alleged’, which is a correct translation-equivalent of Hindi legal term ‘abhikathan’. TermEval failed to label the above term translation as correct due to the fact that its morpholgical form is different to that of the reference term. Here, we show one more example. Consider a source Hindi sentence ‘sbachaav mein koee bhee *gavaah* kee kisee bhee apeel karanevaale ne jaanch nahee kee gaee hai’ and the English reference translation ‘no *witness* in defence has been examined by either of the appellants.’ Here, ‘gavaah’ is a Hindi term and its English equivalent is ‘witness’. The translation of the source sentence by the Hindi-to-English NMT system is ‘no appeal has been examined by any of the *witnesses* in defence’. Here, the translation of the Hindi term ‘gavaah’ is ‘witnesses’ which is correct as per the context of the target translation. Again, TermEval failed to trace this term translation. In Table 11, we report number of instances (i.e. false negatives) in the English–Hindi PB-SMT and NMT tasks, where TermEval failed to label term translations as correct due to the above reason. In Table 11, we see a mix bag of results, i.e. such instances are more seen in PB-SMT when target is Hindi and the other way round (i.e. more seen in NMT) when target is English.

False Negative [Due to Inflectional Issue]			
	PB-SMT	NMT	PB-SMT ∩ NMT
English-to-Hindi	112	87	31
Hindi-to-English	75	107	48

Table 11. False Negatives in the PB-SMT and NMT tasks when TermEval fails to capture a correct term translation whose morphological form is different to the reference term (or one of its LIVs).

We recall the rule that we defined while forming LIV list for a reference term from Section 4.2. *We 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 replaces the reference term in the reference translation, may (syntactically) misfit in the context of the reference translation. This is the reason why the term annotator, at the time of annotation, did not consider such inflectional variation for a reference term. These cases are likely to be seen more with the morphologically-rich languages, e.g. Hindi. However, in our case, we see from Table 11 that this has happened with both Hindi and English. This is to be a challenging problem for any intelligent evaluation metric to address.

In this context, we mention the standard MT evaluation metric METEOR [11] that, to an extent, tackles the above problems (reordering and inflectional issues) with two special modules (i.e. paraphrase and stem matching modules). In future, we intend to investigate this problem in the light of terminology translation in MT.

False Negative [Due to Miscellaneous Reasons]			
	PB-SMT	NMT	PB-SMT ∩ NMT
English-to-Hindi	8	4	-
Hindi-to-English	2	20	-

Table 12. False Negatives in the PB-SMT and NMT translation tasks when TermEval fails to capture correct term translations for various reasons.

Miscellaneous Reasons In Table 12, we report the number of false negatives in PB-SMT and NMT tasks when TermEval fails to capture a correct term translation for various reasons. There are mainly four reasons:

– Reordering and inflectional issue: this can be viewed as the combination the above two types: ‘reordering and inflectional issues’. In short, the translation of a source term contains all words of the reference term (or one of its LIVs) but in a order

different to the reference term (or one of its LIVs) (i.e. 'reordering issue' in above) and one or more words of that translation include inflectional morphemes which are different to those of the words of the reference term (or one of the LIVs) (i.e. 'inflectional issue').
- Term transliteration: translation-equivalent of a source term is the transliteration of the source term itself. We observed that this happened only when the target language is Hindi. In practice, many English terms (transliterated form) are often used in Hindi text (e.g. 'tariff orders', 'exchange control manual').
- Term co-refereed: translation-equivalent of a source term is not found in hypothesis, however, it is rightly co-refereed in target translation.
- Semantically coherence translation: translation-equivalent of a source term is not seen in hypothesis, however, its meaning is rightly transferred into target. For an example, consider the source Hindi sentence "sabhee apeelakartaon ne aparaadh sveekaar nahin kiya aur muqadama chalaaye jaane kee maang kee" and reference sentence "all the appellants pleaded not guilty to the charge and claimed to be tried". 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 source term 'aparaadh sveekaar nahin' is preserved at target translation.

False Negative [Due to Missing LIVs]			
	PB-SMT	NMT	PB-SMT ∩ NMT
English-to-Hindi	33	36	10
Hindi-to-English	24	33	5

Table 13. False negatives in the PB-SMT and NMT tasks when TermEval fails to capture term translations as correct due to the absence of the appropriate LIVs in gold-testset.

Missing LIVs In few cases, we found that a source term is correctly translated into target, but the translation is neither the reference term nor any of its LIVs. These can be viewed as annotation mistake since annotators missed to add relevant LIVs for the reference term into the gold-testset. In Table 13, we report number of false negatives in PB-SMT and NMT tasks when TermEval fails to capture correct term translations due to this reason.

7 Conclusion

In this study, we proposed an automatic evaluation metric for evaluating terminology translation in MT, TermEval. Due to the unavailability of the gold standard for evaluating terminology translation in MT, we adopted a technique that semi-automatically created

a gold-standard test set from English–Hindi judicial domain parallel corpus. We found that TermEval represents a promising metric for the automatic evaluation of terminology translation in MT, while showing very high correlation with the human judgements. We examined why the automatic evaluation technique failed to distinguish term translation in few cases, and identified reasons (e.g. reordering and inflectional issues in term translation) for such aberration.

We also demonstrated our observations on term translation in state-of-the-art PB-SMT and NMT in two evaluation settings: automatic and manual. In manual evaluation, we found that NMT is less error-prone than PB-SMT (0.901 versus 0.917 and 0.870 versus 0.884 ACC (accuracy) scores in English-to-Hindi and Hindi-to-English translation tasks, respectively; differences in scores are statistically significant). We also found that the majority of the term translation errors by the PB-SMT systems are complementary with those by the NMT systems.

In this work, we also created a gold standard test set that can be regarded as an important language resource in MT research. The gold standard test set can also be used for the evaluation of a related natural language processing tasks, e.g. terminology extraction. This can also serve itself as a test-suite for automatic monolingual and bilingual term annotation tasks. We demonstrated various linguistic issues and challenges while creating our gold standard data set, which could provide insights for such annotation scheme.

In future, we aim to make our gold-standard evaluation test set as exhaustive as possible with adding missing LIVs and correcting erroneous LIVs (cf. Section 5.6) of the reference terms. We also intend to incorporate lexical rules in our automatic term evaluation metric, which can help raise its accuracy. We plan to test our evaluation technique 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

1. BitterCorpus. Available online: `https://hlt-mt.fbk.eu/technologies/bittercorpus`, [Accessed on 28-August-2019]
2. Arčan, M., Turchi, M., Tonelli, S., Buitelaar, P.: Enhancing statistical machine translation with bilingual terminology in a cat environment. In: Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014). pp. 54–68 (2014)
3. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. CoRR **abs/1607.06450** (2016), `https://arxiv.org/abs/1607.06450`

4. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations. pp. 1–15. San Diego, CA (2015)
5. Beyer, A.M., Macketanz, V., Burchardt, A., Williams, P.: Can out-of-the-box nmt beat a domain-trained moses on technical data? In: Proceedings of EAMT User Studies and Project/Product Descriptions. pp. 41–46. Prague, Czech Republic (2017)
6. Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., Zeman, D.: Hindencorp - Hindi-English and Hindi-only corpus for machine translation. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC. pp. 3550–3555 (2014), http://www.lrec-conf.org/proceedings/lrec2014/summaries/835.html
7. Burchardt, A., Macketanz, V., Dehdari, J., Heigold, G., Peter, J.T., Williams, P.: A linguistic evaluation of rule-based, phrase-based, and neural mt engines. The Prague Bulletin of Mathematical Linguistics **108**(1), 159–170 (2017), https://content.sciendo.com/view/journals/pralin/108/1/article-p159.xml
8. Cherry, C., Foster, G.: Batch tuning strategies for statistical machine translation. In: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 427–436. Association for Computational Linguistics, Montréal, Canada (2012), http://www.aclweb.org/anthology/N12-1047
9. Cho, K., van Merriënboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Doha, Qatar (October 2014), http://www.aclweb.org/anthology/D14-1179
10. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
11. Denkowski, M., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Sixth Workshop on Statistical Machine Translation. pp. 85–91. Association for Computational Linguistics, Edinburgh, Scotland (July 2011), http://www.aclweb.org/anthology/W11-2107
12. Durrani, N., Schmid, H., Fraser, A.: A joint sequence translation model with integrated reordering. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1045–1054. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/P11-1105
13. Farajian, M.A., Bertoldi, N., Negri, M., Turchi, M., Federico, M.: Evaluation of terminology translation in instance-based neural mt adaptation. In: Proceedings of the 21st Annual Conference of the European Association for Machine Translation. pp. 149–158. Alicante, Spain (2018)
14. Gage, P.: A new algorithm for data compression. C Users Journal **12**(2), 23–38 (1994)
15. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. CoRR **abs/1512.05287** (2016), https://arxiv.org/abs/1512.05287
16. Haque, R., Hasanuzzaman, M., Way, A.: Investigating terminology translation in statistical and neural machine translation: A case study on english-to-hindi and hindi-to-english. In: Proceedings of RANLP 2019: Recent Advances in Natural Language Processing. pp. 437–446. Varna, Bulgaria (2019)
17. Haque, R., Penkale, S., Way, A.: Bilingual termbank creation via log-likelihood comparison and phrase-based statistical machine translation. In: Proceedings of the 4th International Workshop on Computational Terminology (Computerm). pp. 42–51. Dublin, Ireland (2014), http://www.aclweb.org/anthology/W14-4806

18. Haque, R., Penkale, S., Way, A.: TermFinder: log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. Language Resources and Evaluation **52**(2), 365–400 (feb 2018). https://doi.org/10.1007/s10579-018-9412-4
19. 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.Y., Luo, R., Menezes, A., Qin, T., Seide, F., Tan, X., Tian, F., Wu, L., Wu, S., Xia, Y., Zhang, D., Zhang, Z., Zhou, M.: Achieving human parity on automatic Chinese to English news translation. ArXiv e-prints (March 2018)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
21. Heafield, K., Pouzyrevsky, I., Clark, J.H., Koehn, P.: Scalable modified kneser-ney language model estimation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 690–696. Association for Computational Linguistics, Sofia, Bulgaria (August 2013), http://www.aclweb.org/anthology/P13-2121
22. Huang, G., Zhang, J., Zhou, Y., , Zong, C.: A simple, straightforward and effective model for joint bilingual terms detection and word alignment in smt. Natural Language Understanding and Intelligent Applications, ICCPOL/NLPCC 2016 **10102**, 103–115 (2016)
23. Huang, L., Chiang, D.: Forest rescoring: Faster decoding with integrated language models. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 144–151. Association for Computational Linguistics, Prague, Czech Republic (June 2007), http://www.aclweb.org/anthology/P07-1019
24. Junczys-Dowmunt, M., Dwojak, T., Hoang, H.: Is neural machine translation ready for deployment? a case study on 30 translation directions. ArXiv e-prints (2016)
25. Kalchbrenner, N., Blunsom, P.: Recurrent continuous translation models. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1700–1709. Seattle, WA (October 2013), http://www.aclweb.org/anthology/D13-1176
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980
27. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Lin, D., Wu, D. (eds.) Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 388–395. Association for Computational Linguistics, Barcelona, Spain (July 2004), http://acl.ldc.upenn.edu/acl2004/emnlp/pdf/Koehn.pdf
28. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Proceedings of MT Summit X: The Tenth Machine Translation Summit. pp. 79–86. Phuket, Thailand (2005)
29. 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., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: ACL 2007, Proceedings of the Interactive Poster and Demonstration Sessions. pp. 177–180. Prague, Czech Republic (2007)
30. Koehn, P., Och, F.J., Marcu, D.: 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. pp. 48–54. Edmonton, AB (2003)
31. Kunchukuttan, A., Mehta, P., Bhattacharyya, P.: The IIT Bombay English–Hindi parallel corpus. CoRR **1710.02855** (2017), https://arxiv.org/abs/1710.02855
32. Lommel, A.R., Uszkoreit, H., Burchardt, A.: Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics. Tradumática: tecnologies de la traducció **0**(12), 455–463 (12 2014)
33. Macketanz, V., Avramidis, E., Burchardt, A., Helcl, J., Srivastava, A.: Machine translation: Phrase-based, rule-based and neural approaches with linguistic evaluation. Cybernetics and

Information Technologies **17**(2), 28–43 (2017), https://content.sciendo.com/view/journals/pralin/108/1/article-p159.xml
34. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1), 19–51 (2003)
35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. ACL, Philadelphia, PA (2002)
36. Pazienza, M.T., Pennacchiotti, M., Zanzotto, F.M.: Terminology extraction: an analysis of linguistic and statistical approaches. In: Sirmakessis, S. (ed.) Knowledge Mining, pp. 255–279. Springer (2005)
37. Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostay, T.: Term extraction, tagging, and mapping tools for under-resourced languages. In: Proceedings of the 10th Conference on Terminology and Knowledge Engineering (TKE 2012). pp. 193–208. Madrid, Spain (06 2012)
38. Popović, M.: chrf: character n-gram f-score for automatic mt evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. Association for Computational Linguistics, Lisbon, Portugal (September 2015), http://aclweb.org/anthology/W15-3049
39. Popović, M.: Comparing language related issues for nmt and pbmt between German and English. The Prague Bulletin of Mathematical Linguistics **108**(1), 209–220 (2017)
40. Press, O., Wolf, L.: Using the output embedding to improve language models. CoRR **abs/1608.05859** (2016), http://arxiv.org/abs/1608.05859
41. Sennrich, R., Haddow, B., Birch, A.: Improving neural machine translation models with monolingual data. CoRR **abs/1511.06709** (2015), http://arxiv.org/abs/1511.06709
42. Sennrich, R., Haddow, B., Birch, A.: 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). pp. 1715–1725. Association for Computational Linguistics, Berlin, Germany (August 2016), http://www.aclweb.org/anthology/P16-1162
43. Skadiņš, R., Puriņš, M., Skadiņa, I., Vasiļjevs, A.: Evaluation of SMT in localization to under-resourced inflected language. In: Proceedings of the 15th International Conference of the European Association for Machine Translation (EAMT 2011). pp. 35–40. Leuven, Belgium (2011)
44. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: In Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006). pp. 223–231. Cambridge, Massachusetts (2006)
45. Specia, L., Harris, K., Blain, F., Burchardt, A., Macketanz, V., Skadiņa, I., Negri, M., , Turchi, M.: Translation quality and productivity: A study on rich morphology languages. In: Proceedings of MT Summit XVI, the 16th Machine Translation Summit. pp. 55–71. Asia-Pacific Association for Machine Translation, Nagoya, Japan (2017)
46. Stanojević, M., Sima'an, K.: Beer: Better evaluation as ranking. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. pp. 414–419. Association for Computational Linguistics, Baltimore, Maryland, USA (June 2014), http://www.aclweb.org/anthology/W/W14/W14-3354
47. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. pp. 3104–3112. NIPS'14, Montreal, Canada (2014)
48. Tiedemann, J.: Parallel data, tools and interfaces in OPUS. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012). pp. 2214–2218. Istanbul, Turkey (2012), http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

49. Toral, A., Sánchez-Cartagena, V.M.: A multifaceted evaluation of neural versus phrase-based machine translation for 9 language directions. CoRR **abs/1701.02901** (2017), http://arxiv.org/abs/1701.02901
50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. CoRR **abs/1706.03762** (2017), http://arxiv.org/abs/1706.03762
51. Vaswani, A., Zhao, Y., Fossum, V., Chiang, D.: Decoding with large-scale neural language models improves translation. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1387–1392. Association for Computational Linguistics, Seattle, Washington, USA (October 2013), http://www.aclweb.org/anthology/D13-1140
52. Vintar, v.: Terminology translation accuracy in statistical versus neural mt: An evaluation for the English–Slovene language pair. In: Du, J., Arcan, M., Liu, Q., Isahara, H. (eds.) Proceedings of the LREC 2018 Workshop MLP–MomenT: The Second Workshop on Multi-Language Processing in a Globalising World and The First Workshop on Multilingualism at the intersection of Knowledge Bases and Machine Translation. pp. 34–37. European Language Resources Association (ELRA), Miyazaki, Japan (may 2018)
53. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016), http://arxiv.org/abs/1609.08144
54. Yeh, A.: More accurate tests for the statistical significance of result differences. In: Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING 2000. pp. 947–953. Saarbrücken, Germany (2000)
55. Ziemski, M., Junczys-Dowmunt, M., Pouliquen, B.: The united nations parallel corpus v1.0. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Portorož, Slovenia (2016)