information

Article

Terminology Translation in Low-Resource Scenarios

Rejwanul Haque [1,*], Mohammed Hasanuzzaman [2] and Andy Way [1]

[1] School of Computing, Dublin City University, Dublin 9 Glasnevin, Ireland
[2] Department of Computer Science, Cork Institute of Technology, T12 P928 Cork, Ireland
* Correspondence: rejwanul.haque@adaptcentre.ie; Tel.: +353-1-7005-074

Received: 30 June 2019; Accepted: 22 August 2019; Published: 30 August 2019

Abstract: Term translation quality in machine translation (MT), which is usually measured by domain experts, is a time-consuming and expensive task. In fact, this is unimaginable in an industrial setting where customised MT systems often need to be updated for many reasons (e.g., availability of new training data, leading MT techniques). To the best of our knowledge, as of yet, there is no publicly-available solution to evaluate terminology translation in MT automatically. Hence, there is a genuine need to have a faster and less-expensive solution to this problem, which could help end-users to identify term translation problems in MT instantly. This study presents a faster and less expensive strategy for evaluating terminology translation in MT. High correlations of our evaluation results with human judgements demonstrate the effectiveness of the proposed solution. The paper also introduces a classification framework, *TermCat*, that can automatically classify term translation-related errors and expose specific problems in relation to terminology translation in MT. We carried out our experiments with a low resource language pair, English–Hindi, and found that our classifier, whose accuracy varies across the translation directions, error classes, the morphological nature of the languages, and MT models, generally performs competently in the terminology translation classification task.

Keywords: machine translation; terminology translation; phrase-based statistical machine translation; neural machine translation; terminology 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, the words "terminal" ("a bus terminal" or "computer terminal") and "play" ("play music" or "play football") 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"). When encountering a legal document, the translation of "charge" has to be the particular Hindi word "aarop". The target translation could lose its meaning if the term translation and domain knowledge is not taken into account. Accordingly, the preservation of domain knowledge from source to target is pivotal in any translation workflow (TW), and this is one of the customer's primary concerns in the translation industry, especially for critical domains such as medical, transportation, military, legal and aerospace. 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 areas in MT research. No standard automatic MT evaluation metric (e.g., BLEU [1]) can provide much information on how good or bad an MT system is at translating domain-specific expressions. To the best of our knowledge, as of now, no one has proposed any effective way to evaluate terminology translation in MT automatically. In industrial TWs, TSPs generally hire human experts related to the concerned domain for identifying term translation problems in MT. Nevertheless, such a process is expensive and time consuming. Moreover, in an industrial setting, retraining of customer-specific MT engines from scratch is carried out quite often when a reasonable amount of new training data pertaining to the domain and style on which that MT system was built or a new state-of-the-art MT technique are available. In industry, carrying out human evaluation on term translation each time from scratch when an MT system is updated would be exorbitant in a commercial context. This is an acute problem in industrial TW, which TSPs are desperate to solve. A suitable solution to the problem of terminology translation evaluation would certainly aid MT users who want to assess their MT systems quickly in the area of domain-specific term translation.

This work presents a faster and less-expensive evaluation strategy [2] that can help quickly assess terminology translation quality in automatic translation. We demonstrate a semi-automatic terminology annotation strategy from which a gold standard for evaluating terminology translation in automatic translation can be created. We use our in-house bilingual term annotation tool, *TermMarker*, for the annotation process. In short, TermMarker marks source and target terms on either side of a test set, incorporating lexical and inflectional variations of the terms relevant to the context in which they appear, by exploiting the automatic terminology extraction technique of [3,4]. The annotation technique needs little manual intervention to validate the term tagging and mapping in the annotation interface. In an industrial setup, TSPs would view this method as an ideal and one-time solution, since the annotation scheme is a less expensive and faster exercise and will result in a reusable gold standard for measuring the MT system's term translation quality. In this study, we create a gold standard test set from a legal domain dataset (i.e., judicial proceedings). From now, we call the gold standard evaluation test set the *gold-test set*.

We introduce an automatic evaluation metric, *TermEval*, to quickly assess terminology translation quality in automatic translation. Going a step further, we propose an automatic classification model, *TermCat*, that can automatically categorise erroneous term translations in MT. TermCat can provide the MT users with more specific information on the nature of terminological errors that an MT system can commit. Both TermEval and TermCat work on the top of the gold standard dataset (cf. Section 4).

To check how TermEval correlates with human judgement, the translation of each source term (of the gold standard test set) is validated by human evaluators. At the same time, the evaluators classify erroneous terminology translations into a set of predefined categories. This way, we create a reference dataset for evaluating TermEval and TermCat. Our evaluation results show that TermEval represents a promising metric for evaluating terminology translation in MT, as it shows very high correlations with the human judgements. We also found that TermCat, whose accuracy varies through various factors (e.g., translation directions, error classes, the morphological nature of the languages and MT models), generally performs competently in the terminology translation error classification task.

Hindi is one of the official languages of India and spoken by 322 million peoples worldwide. Hindi is written in Devanagari script [5]. Like other Indian languages, Hindi is also a free word order (used with emphasis and complex structures) language. As far as building a good quality NMT engine is concerned, the standard practice requires a large parallel corpus (i.e. an order of over 10 millions segments) [6]. The publicly available English-to-Hindi parallel corpus contains just over one million segments [7] (see Section 3.3). Moreover, this corpus is not made of a single domain, i.e. it is a mixture of various domains. Data scarcity is even more prominent for Hindi and other non-English languages, i.e. many translation pairs involving Hindi and non-English languages have either no data or tiny-sized data. There is a genuine need for creating parallel corpora for Indic languages including Hindi as far as the research in MT and related NLP fields is concerned. The MT research community see all Indic languages including Hindi as the resource-poor languages.

Phrase-based statistical MT (PB-SMT) [8], a predictive modelling approach to MT, was the main paradigm in MT research for more than two decades. Neural MT [9–13], 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 five years, there has been incremental progress in the field of NMT [12,13] to the point where some researchers are claiming parity with human translation [14]. Currently, NMT is regarded as a preferred alternative to PB-SMT [8] 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. 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.

To summarize, our main contributions in this article are as follows:

1. We present a faster and less expensive annotation scheme that can semi-automatically create a reusable gold standard evaluation test set for evaluating terminology translation in MT. This strategy provides a faster and less expensive solution compared to a slow and expensive manual evaluation process.
2. We highlight various linguistic phenomena in relation to the annotation process on English and a low resource language, Hindi.
3. We present an automatic metric for evaluating terminology translation in MT, namely TermEval. We also demonstrate a classification framework, TermCat, that can automatically classify terminology translation-related errors in MT. TermEval is shown to be a promising metric, as it shows very high correlations with the human judgements. TermCat achieves competitive performance in the terminology translation error classification task.
4. We compare PB-SMT and NMT on terminology translation in two translation directions: English-to-Hindi and Hindi-to-English. We present the challenges in relation to the investigation of the automation of the terminology translation evaluation and error classification processes in the low resource scenario.

The remainder of the paper is organised as follows. In Section 2, we discuss related work. Section 3 describes the MT systems used in our experiments. In Section 4, we present how we created a gold standard dataset and examine challenges in relation to the termbank creation process. Sections 5 and 6 present our evaluation metric, TermEval, and classification model, TermCat, respectively. Section 7 reports our evaluation plan and experimental results, including discussion and analysis, while Section 8 concludes and provides avenues for further work.

2. Related Work

Multiword units (MWUs) are overwhelmingly present in terminology [15]. The book *Multiword units in machine translation and translation technology* [16] explores the computational treatments of how MWUs (e.g. idioms, collocations, terminology) can be handled in NLP, particularly in MT and translation technology [17]. Since we takes terminology into consideration for our investigation, the literature survey section talks about those works that studied handling techniques of single-word and multiword terms in NLP tasks relevant to ours. Hence, this section is divided into three subsections: *Terminology Annotation*, *Terminology Translation Evaluation Methods* and *Terminology Translation in PB-SMT & NMT*.

2.1. Terminology Annotation

Annotation techniques have been widely studied in many areas of natural language processing (NLP). However, terminology annotation is a rarely investigated domain in NLP due to many challenges [18]. Pinnis et al. [19] investigated term extraction, tagging and mapping techniques for under-resourced languages. They mainly presented 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 primarily focused on acquiring bilingual terms from comparable web-crawled narrow domain corpora. We refer interested readers to Section 2.3 of [18], which presents studies relating to annotation and evaluation of multilingual automatic terminology extraction from comparable corpora.

In order to evaluate the quality of the bilingual terms in MT, Arčan et al. [20] 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 [21]. 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 [22]. 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 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 [3,4]. 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 [23]). 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.

2.2. Terminology Translation Evaluation Methods

Farajian et al. [24] 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. Provided with a termbank, this metric looks for source and target terms in the reference set and translated documents. There is a potential problem with this evaluation strategy. It might be the case that 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 a case, the evaluation metric would make a hit count, which is not correct. In addition to the above aberration, there are two more issues that Farajian et al. [24] could not address, namely the problem relating to the translation of the ambiguous terms (cf. Section 4.4) and the consideration of lexical and inflectional variations for a reference term (cf. Section 4.2).

Automatic error analysis methods [25,26] can expose specific problems in translation with an MT system. Popović and Ney [25] proposed a framework for automatic error analysis and classification of translations, which essentially provides more specific information of certain problems in translations. In their work, they focused on five basic error categories: (a) inflectional error, (b) reordering error, (c) extra word, (d) missing words and (e) incorrect lexical choice. More recently, in order to compare the performance of NMT and classical MT paradigms, Bentivogli et al. [26] automatically classified translational errors using a coarse-grained error typology: (a) morphology errors, (b) lexical errors and (c) word order errors. To the best of our knowledge, no one has tried to classify terminological errors in MT automatically. We demonstrate a classification framework, TermCat, that can automatically classify terminology translation-related errors and provide deeper knowledge on the nature of terminological errors.

2.3. Terminology Translation in PB-SMT & NMT

This section presents papers that have looked into terminology translation in MT. Burchardt et al. [27] conducted 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, despite obtaining the lowest average score, PB-SMT was the best-performing system on named entities and terminology translation. However, when tested in the reverse translation direction (i.e., English-to-German), a commercial NMT engine becomes the winner as far as term translation is concerned. In a similar experimental setup, Macketanz et al. [28] reported that their PB-SMT system outperformed NMT on terminology translation on both in-domain (IT domain) and general domain test suites in an English-to-German translation task. Specia et al. [29] carried out an error annotation process using the multidimensional quality metrics (MQM) error annotation framework [30] in an MT post-editing environment. The list of errors was divided into three main categories: accuracy, fluency and terminology. According to the annotation results, more terminology-related errors were found in NMT translations than in PB-SMT translations in the English-to-German task (139 vs. 82), and the other way round in the English-to-Latvian task (31 vs. 34). From their manual evaluation procedure, Beyer et al. [31] reported that PB-SMT outperformed NMT on term translation, which they speculated could be because their technical termbank was part of the training data used for building their PB-SMT system. Vintar [32] conducted an automatic and small-scale human evaluation on the terminology translation quality of the Google Translate NMT model [33] compared to its earlier PB-SMT model for the Slovene–English language pair and in the specialised domain of karstology [34]. The evaluation result on the Slovene-to-English task confirmed NMT to be slightly better than PB-SMT in terminology translation, while the opposite direction (i.e., English-to-Slovene task) showed the reversed picture with PB-SMT outperforming NMT. Vintar [32] carried out a small qualitative analysis, by counting terms that are dropped in target translations and detailing instances where MT systems often failed to preserve domain knowledge. Recently, we conducted comparative qualitative evaluation and comprehensive error analysis on terminology translation in PB-SMT and NMT, and demonstrated a number of important findings [35], e.g. NMT systems omit more terms in translation than PB-SMT systems, majority of the errors made by PB-SMT systems are complementary to those made by NMT systems. As far as terminology translation quality evaluation in PB-SMT alone is concerned, given its relatively longer history, there have been a number of papers on that topic. For example, Huang et al. [36] investigated term translation in a PB-SMT task and observed that more than 10% of high-frequency terms were incorrectly translated by their PB-SMT decoder, although the system's BLEU score was quite high, i.e., 63.0 BLEU. One common event in the above papers [27–29,31,32,35,36] was that the authors carried out a human evaluation in order to measure terminology translation quality in MT, which, as mentioned earlier, is subjective and a slow and expensive process. In this paper, we present a faster and less-expensive evaluation strategy for preparing a gold standard, based on which our metric, TermEval, can measure terminology translation quality in MT. Furthermore, we demonstrate a classification framework that can expose the nature of terminological errors in MT.

3. MT Systems

3.1. PB-SMT System

To build our PB-SMT systems, we used the Moses toolkit [37]. We used a five-gram LM trained with modified Kneser–Ney smoothing [38] using the KenLM toolkit [39]. For LM training, we combined a large monolingual corpus with the target-side of the parallel training corpus. Additionally, we trained a neural LM with the NPLM toolkit [40] 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 lexicalized reordering probabilities (*wbe-mslr-bidirectional-fe-allff*); (c) 2 5-gram LM probabilities (Kneser–Ney and NPLM); (d) 5 OSM features [41] and (e) word-count and distortion penalties. In our experiments, word alignment models were trained using the GIZA++ toolkit [42], phrases were extracted following the *grow-diag-final-and* algorithm of [8], Kneser–Ney smoothing was applied at the level of phrase scoring, and a smoothing constant (0.8u) was used for training lexicalized reordering models. The weights of the parameters were optimized using the margin-infused relaxed algorithm [43] on the development set. For decoding, the cube-pruning algorithm [44] was applied, with a distortion limit of 12. We called the English-to-Hindi and Hindi-to-English PB-SMT systems EHPS and HEPS, respectively.

3.2. NMT System

To build our NMT systems, we used the MarianNMT [45] toolkit. The NMT systems are Google Transformer models [13]. In our experiments, we followed the recommended best setup from [13]. The tokens of the training, evaluation and validation sets were segmented into sub-word units using the byte-pair encoding (BPE) technique [46] proposed by [47]. Since English and Hindi are written in Roman and Devanagari scripts and have no overlapping characters, BPE was applied individually on the source and target languages. We performed 32,000 join operations. Our training setup was as follows. We considered the size of the encoder and decoder layers to be six. As in [13], we employed residual connection around layers [48], followed by layer normalisation [49]. The weight matrix between embedding layers was shared, similar to [50]. Dropout [51] between layers was set to 0.10. We used mini-batches of a size of 64 for updating. The models were trained with the Adam optimizer [52], with the learning-rate set to 0.0003 and reshuffling the training corpora for each epoch. As in [13], we also used the learning rate warm-up strategy for Adam. The validation on development set was performed using three cost functions: cross-entropy, perplexity and BLEU. The early stopping criterion was based on cross-entropy; however, the final NMT systems were selected as per the highest BLEU scores on the validation set. The beam size for search was set to 12.

Initially, we used a parallel training corpus to build our English-to-Hindi and Hindi-to-English baseline transformer models. We translated monolingual sentences (cf. Table 1) with the baseline models and created source synthetic sentences [53,54]. Then, we append this synthetic training data to the parallel training data and retrained the baseline models. We made our final NMT model with ensembles of four models (top four models as per the BLEU scores on the validation set) that were sampled from the training run. We called 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 [7], which is compiled from a variety of existing sources, e.g., OPUS [21]. That is why the parallel corpus is a mixture of various domains. For building additional language models (LMs) for Hindi and English, we used the HindEnCorp monolingual corpus [55] and monolingual data from various sources (e.g., the European Parliamentary proceedings [56]) from the OPUS project, respectively. The corpus statistics are shown in Table 1. We selected 2000 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 judgements). 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 tokenization for English and Hindi, we used the standard tokenization tool [57] of the Moses toolkit.

Table 1. Corpus statistics.

English–Hindi Parallel Corpus			
	Sentences	**Words (English)**	**Words (Hindi)**
Training set	1,243,024	17,485,320	18,744,496
(Vocabulary)		180,807	309,879
Judicial	7374	179,503	193,729
Development set	996	19,868	20,634
Test set	2000	39,627	41,249
Monolingual Corpus		**Sentences**	**Words**
Used for PB-SMT Language Model			
English		11 M	222 M
Hindi		10.4 M	199 M
Used for NMT Back Translation			
English		1 M	20.2 M
Hindi		903 K	14.2 M

3.4. PB-SMT versus NMT

In this section, we present the comparative performance of the PB-SMT and NMT systems using a range of automatic evaluation metrics: BLEU, METEOR [58] and TER [59]. BLEU, METEOR and TER are standard metrics that are widely used by the MT community. Note that TER is an error metric, which means lower values indicate better translation quality. We report evaluation results in Table 2. Additionally, we performed statistical significance tests using bootstrap resampling methods [60]. The confidence level (%) of the improvement obtained by one MT system with respect to the another MT system is reported. An improvement in system performance at a confidence level above 95% was assumed to be statistically significant.

Table 2. Performance of PB-SMT and NMT systems on automatic evaluation metrics. EHPS, English-to-Hindi system; EHNS, English-to-Hindi system; HEPS, Hindi-to-English PB-SMT system; HENS, Hindi-to-English NMT system.

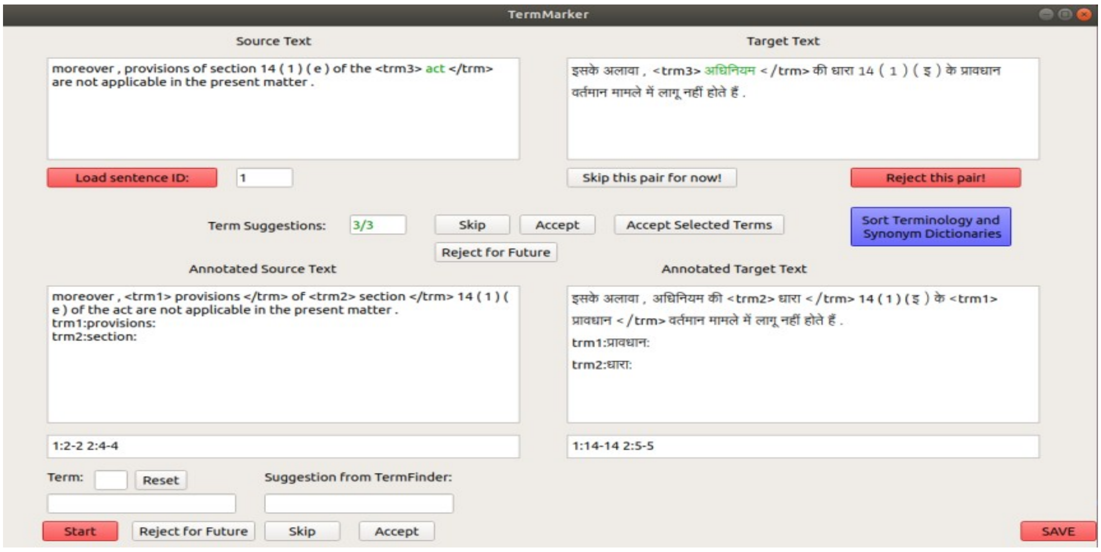
	BLEU	**METEOR**	**TER**
EHPS	28.8	30.2	53.4
EHNS	36.6 (99.9%)	33.5 (99.9%)	46.3 (99.9%)
HEPS	34.1	36.6	50.0
HENS	39.9 (99.9%)	38.5 (98.6%)	42.0 (99.9%)

As can be seen from Table 2, EHPS and EHNS produced 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, underestimated the translation quality, given the relatively free word order in Hindi, given that we had just a single reference translation set for evaluation. Many TSPs consider a 30.0 BLEU score as a benchmarking value and use those MT systems in their TW that produce BLEU scores above this score. For example, the work in [61] successfully used an English-to-Latvian MT system with a similar BLEU score (35.0 BLEU) in the SDL Trados CAT tool [62]. In this perspective, EHPS was just below par, and EHNS was well above the benchmarking value.

As far as the Hindi-to-English translation task was concerned, HEPS and HENS produced reasonable BLEU scores (34.1 BLEU and 39.9 BLEU) on the test set. As expected, translation quality from the morphologically-rich to morphologically-poor language improved. The differences in BLEU scores of PB-SMT and NMT systems in both the English-to-Hindi and Hindi-to-English translation tasks were statistically significant. This trend was observed with the other evaluation metrics.

4. Creating Gold Standard Evaluation Set

This section presents our annotation scheme. For evaluating terminology translation with our MT systems, we used the test set (cf. Table 1) that contained 2000 sentence-pairs from the judicial domain. We annotated the test set by marking term-pairs on the source and target sides of the test set for creating a gold standard evaluation set. The annotation process was accomplished with our own bilingual term annotation tool, *TermMarker*, a user-friendly GUI developed with PyQT5 [63]. Figure 1 shows a screenshot of TermMarker's interface. The annotation process starts by displaying a source–target sentence-pair from the test set in the 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, and the source–target term-pair is marked. The annotation process is simple, and is carried out manually. The annotators (native Hindi evaluators with excellent English skills) were instructed to mark those words as terms that belong to the legal or judicial domains. The annotators were also instructed to mark those sentence-pairs from the test set that contained 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 called the remaining 1270 sentence-pairs our *gold-test set*. Each sentence-pair of the gold-test set contained at least one aligned source–target term-pair. The gold-test set is publicly available to the research community via a website [64].



moreover , provisions of section 14 (1) (e) of the <trm3> act </trm> are not applicable in the present matter .

इसके अलावा , <trm3> अधिनियम < /trm> की धारा 14 (1) (ड) के प्रावधान वर्तमान मामले में लागू नहीं होते हैं .

1

3/3

moreover , <trm1> provisions </trm> of <trm2> section </trm> 14 (1) (e) of the act are not applicable in the present matter .
trm1:provisions:
trm2:section:

इसके अलावा , अधिनियम की <trm2> धारा < /trm> 14 (1) (ड) के <trm1> प्रावधान < /trm> वर्तमान मामले में लागू नहीं होते हैं .
trm1:प्रावधान:
trm2:धारा:

1:2-2 2:4-4

1:14-14 2:5-5

Figure 1. TermMarker.

4.1. Annotation Suggestions from Bilingual Terminology

TermMarker supports annotation suggestions from an externally-supplied terminology. We recommend this option for faster annotation, although it 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 [3,4], which is regarded as the state-of-the-art terminology extraction technique and works well even on as few as 5000 parallel segments. As can be seen from Figure 1, given a source-target (English–Hindi) legal domain sentence pair, TermMarker suggests an annotation by tagging the legal term "act" in the English sentence and its target translation (i.e., Hindi term) "adhiniyam" in the Hindi sentence. 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 is to be used in the annotation process.

In Table 3, we show the statistics of the occurrences of terms in the gold standard evaluation set (i.e., gold-test set). We found 3064 English terms and their target equivalents (3064 Hindi terms) in the source and target sides of gold-test set, respectively. We observed the presence of nested terms (i.e., overlapping terms) in the gold-test set, 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 can only have one superterm. TermMarker allows us to annotate both subterms and superterms.

Table 3. Statistics of the occurrences of terms in the gold-test set.

Number of Source—Target Term-Pairs		3064
English	Terms with LIVs LIVs/Term	2057 5.2
Hindi	Terms with LIVs LIVs/Term	2709 8.4

4.2. Variations of Terms

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, the translation of the English word "affidavit" has multiple target equivalents (LIVs) 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 the 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, the 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 (lexical or inflectional) variations 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 alternative terms. We make the list of lexical and inflectional variations (LIVs) for the term as exhaustive as possible. In Table 3, we report the number of English and Hindi terms for which we added lexical or inflectional variations and the average number of variations per such term. As expected, both the numbers were higher in Hindi than English.

During annotation, the user could manually add relevant variations for a term through TermMarker's interface. We again exploited the method of [3,4] for obtaining variation suggestions for a term. The automatically-extracted bilingual terminology of [3,4] 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 alternative suggestions of the target term. As in 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 also has an option to conduct annotation in both directions (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 solve this inconsistency problem, TermMarker includes a *check-up module* that traces a term's annotation history, mainly by storing rejected and skipped items. The check-up module notifies the human annotators when any of the preceding sentences has to be annotated for a newly-included term-pair or variation of a term in the termbank.

Two annotators took part in the annotation task, and two sets of annotated data were obtained. The term-pairs of the gold-test set were finalised on the basis of the annotation agreement by the two annotators, i.e., we kept those source–target term-pairs in the gold-test set for which both annotators agreed that the source and target entities were terms and aligned. On completion of the annotation process, inter-annotator agreement was computed using Cohen's kappa [65] at the word level. This means for a multiword term, we considered the number of words in it for this calculation. For each word, we counted an agreement whenever both annotators agreed that it was a term (or part of a term) or a non-term entity. We found the kappa coefficient to be very high (i.e., 0.95) for the annotation task. This indicated that our terminology annotation was of excellent quality. The final LIV list for a term was the union of the LIV lists created by the annotators.

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 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 the 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., lexicon) for the same reason. Hence, the terminology annotation task is inseparably associated with the term's degree of ambiguity. This hypothesis is unquestionably true for the type of data domain (e.g., judicial) we considered in this work.

4.5. Measuring the Performance of TermMarker

We tested whether the use of annotation and variation suggestions from bilingual terminology [3,4] speeds up the annotation process. For this, we divided our judicial domain test set (2000 segments, cf. Table 1) into 10 sets, each of which contained 200 segment-pairs. Note that we started our annotation process on Set 1 and ended on Set 10. We counted 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 2. The x-axis and y-axis of Figure 2 represent the segment set number and acceptance ratio, respectively. We see from Figure 2 that both curves (annotation and variation suggestions) moved upward over time, i.e., the acceptance rate of suggestions increased over time. This is because anomalous entries were rejected from the terminology, and new valid entries were added into the terminology all the time, which made the annotation process iteratively faster.

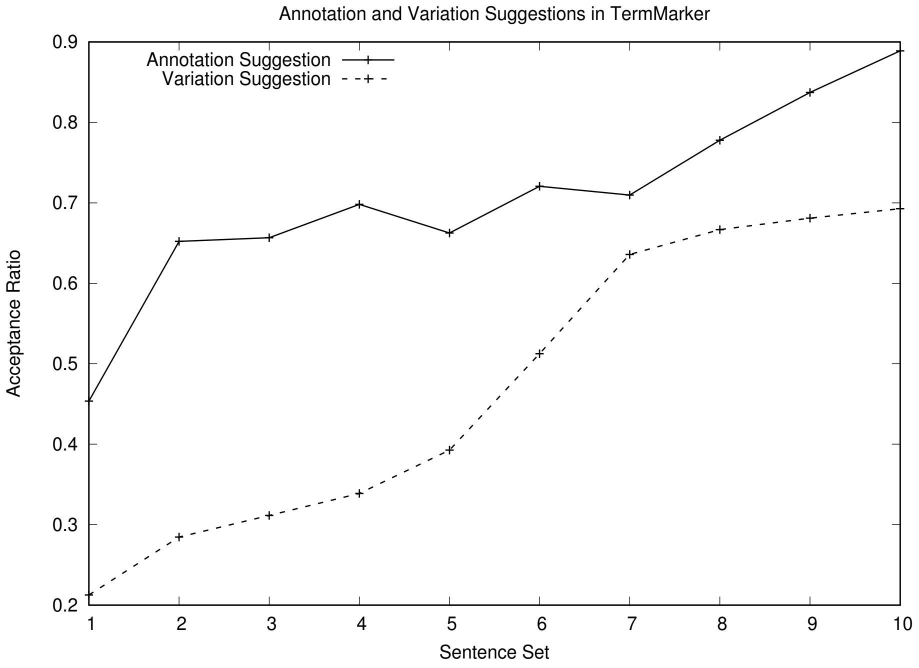


Figure 2. Curves for acceptance ratio of suggestions.

5. TermEval

This section explains our proposed evaluation metric, TermEval. The metric starts the evaluation process by forming a set of tuples with each source sentence from the test set and its translation (i.e., hypothesis). The tuples were formed with the list of source terms appearing in the source sentence, their reference translations (i.e., terms) and LIVs (lexical and inflection variations) of the reference terms. First, we looked for the reference terms (or the LIVs of the reference terms) in the hypothesis. We employed the *longest reference term first* strategy in the search. If a reference term (or any of its LIVs) of a source term was found in the hypothesis, this indicated that the MT system had correctly translated the source term into the target language. If this was not the case, there was likely to be an error in that term translation. At the end of the iteration (over all tuples), the next sentence from the test set was taken into consideration for calculation. Finally, we obtained the TermEval score over the test set. The evaluation method included an additional procedure that took nested overlapping terms and multiple identical reference terms (or variations) into account. More formally, TermEval was calculated given the test set and translation set using (1):

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

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

We obtained TermEval scores to evaluate the performance of our MT systems (cf. Section 3) on the gold-test set (cf. Section 4), which are reported in Section 7.3. We also tested how TermEval correlates with human judgements and report our findings there. In this context, 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 a 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 a dependency.

6. TermCat

This section presents our proposed classification framework, TermCat, which can automatically classify terminology translation errors in MT. First, we discuss the error types we considered for the classification task (cf. Section 6.1), and then, we describe how our classification framework TermCat operates (cf. Section 6.2).

6.1. Error Classes

We translated the test set sentences with our MT systems (cf. Section 3) and sampled three hundred translations from the whole translation set. We then manually inspected the terminology translations, noting especially the patterns of the term translation-related errors. From our observations, we found that the terminology translation-related errors can roughly be classified into five primary categories. As a consequence, we made a high-level classification of the terminology translation-related errors in MT, as follows:

1. Reorder error (RE): the translation of a source term forms the wrong word order in the target language.
2. Inflectional Error (IE): the translation of a source term inflicts a morphological error (e.g., includes an inflectional morpheme that is a misfit in the context of the target translation, which essentially causes a grammatical error in translation).
3. Partial error (PE): the MT system correctly translates part of a source term into the target language and commits an error for the remainder of the source term.
4. Incorrect lexical selection (ILS): the translation of a source term is an incorrect lexical choice.
5. Term drop (TD): the MT system omits the source term in translation.

6.2. Classification Framework

We detail how TermCat classifies incorrect term translations into a set of fine-grained error categories below. We refer back to Section 5 where we described how TermEval works. In short, for a given source sentence and its translation (i.e., hypothesis), TermEval forms a set of tuples and looks for the reference translations in the hypothesis. At the end of the iteration (over all tuples), any remaining unlabelled tuples are passed to TermCat. Then, TermCat invokes its error identification modules that classify terminological errors into a set of categories. Thus, TermCat works on the top of TermEval. TermCat is a generic model, but makes use of language-specific linguistic processors and lexical knowledge bases. TermCat is a rule-based application; a flowchart with its error identification modules is illustrated in Figure 3. We explain each of the TermCat's error identification modules below.

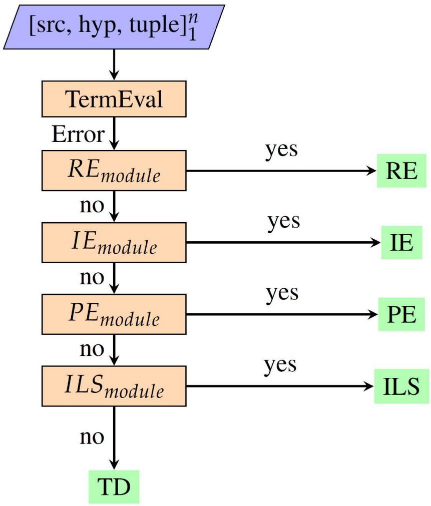


Figure 3. Flowchart of TermCat.

6.2.1. Reorder Error Identification Module

TermCat's first error identification module involves the reordering problem in terminology translation (cf. Figure 3). The RE module starts working once TermEval ends processing a source sentence and its translation. The RE module looks for multi-word reference terms (or multi-word LIVs of the reference terms) in the hypothesis. If words of a multi-word reference term (or one of the multi-word LIVs of a reference term) appear in the hypothesis, but in a different order, then this indicates that a reordering error has occurred in translation, and TermCat labels the corresponding source term and its translation with RE. At the end of the iteration, TermCat invokes the next module (IE identification), passing remaining unlabelled tuples to subsequent modules.

6.2.2. Inflectional Error Identification Module

This module is responsible for tracing inflectional errors in terminology translation. In order to do this, the IE module employs a stem-level comparison of the reference terms (or their LIVs) with the words of the hypothesis. In order to obtain stems from words, we used the Porter stemmer [66] for English and a lightweight stemmer [67] for Hindi. In case of a successful search, the classifier labels the corresponding source term and its translation with IE. At the end of the iteration, TermCat invokes a sub-procedure for processing any remaining unlabelled tuples. Note that this module also can trace those term translation-related errors that are a combination of both reordering and inflectional errors, i.e., when the translation of a source term makes an inflectional error, as well as the wrong word order. The sub-procedure is identical to that of the RE identification module, barring the comparison, which is performed here at the stem level. In the case of a successful search, TermCat labels the corresponding source term and its translation with IE. At the end of the iteration, the classifier invokes the next module (PE identification) for processing any remaining unlabelled tuples.

6.2.3. Partial Error Identification Module

The PE (partial error) module spots those source terms whose target translations are partially correct. This module looks for a part of the reference terms (or its variations) in the hypothesis. In the case of a successful search, TermCat labels the corresponding source term and its translation with PE since there is likely to be an error with the remaining part of the source term. Note that stop words (e.g., prepositions, articles) that the reference terms or its variations include are not considered during the search. Consider a multi-word legal term "acts of negligence", which contains a stop word "of". We cannot look for "of" in the hypothesis for spotting the term's partial error since "of" commonly appears in English sentences. At the end of the iteration, TermCat invokes the next module (ILS) for processing any remaining unlabelled tuples.

6.2.4. Incorrect Lexical Selection Identification Module

This module identifies term's incorrect lexical choice. We describe how it operates below. First, TermCat collects all relevant translation-equivalents for each of the source terms. This is accomplished in two different ways: (i) for a single word reference term, we obtained its synonyms from WordNet [68] for English and Hindi WordNet [69] for Hindi, and (ii) for a single or multi-word source term, we obtained translation options from the PB-SMT phrase table. For each source term, we excluded its reference term and the LIVs of the reference term from the list of collected relevant translation-equivalents. Then, ILS Module looks for the remaining relevant translation-equivalents of a source term in the hypothesis. A successful search indicates that the MT system has translated the corresponding source term into the target language and the translation is an incorrect lexical choice, and TermCat labels this source term and its translation with ILS. At the end of the iteration, TermCat calls a sub-procedure for processing any remaining unlabelled tuples. It could be a case that translation of a part of a source term is an incorrect lexical choice, and for the remaining part of the source term, the MT system commits another error. The sub-procedure is a clone of the main identification module barring an exception, i.e., it looks for a part of the target-equivalents of the source terms in the hypothesis. A successful search indicates that the MT system has partially translated the source term into the target language, and the translation is an incorrect lexical choice. Hence, TermCat labels the corresponding source term and its translation with ILS. Like PE, in this sub-procedure, stop words that the reference terms include are not considered during the search.

Since for any remaining unlabelled tuples, no translation-equivalents (term and non-term entities) of the source terms either partially or as a whole were found in the hypothesis, we concluded that the MT system had omitted these source terms in translation. Hence, TermCat labels them with TD (term drop).

The above describes the workings of TermCat. Given the list of terms from a source text and their reference translations, TermCat can automatically classify terminology translation-related errors in automatic translation. Hence, our gold-test set (cf. Section 4) is to be seen as an ideal test set for the classification task. We tested TermCat on the gold-test set in four translation tasks, i.e., the English-to-Hindi and Hindi-to-English PB-SMT and NMT tasks, and report its performance in the terminology translation error classification task in Section 7.5.

7. Results, Discussion and Analysis

This section first describes our manual evaluation plan and results. Then, we test our proposed metric, TermEval, and classification model, TermCat, and discuss the results.

7.1. Manual Evaluation

This section presents our manual evaluation plan. As discussed in Section 6.1, we divided terminology translation-related errors into five main categories (reorder errors (RE), inflectional errors (IE), partial errors (PE), incorrect lexical selection (ILS) and term drop (TD)). In addition to these five primary error classes, we defined another primary error class REM (i.e., remaining). If there was an error in relation to the translation of a source term, whose category was beyond the first five error categories, we placed that term translation under the REM category.

As mentioned above, the sentences of gold-test set were translated with the English-to-Hindi and Hindi-to-English MT systems (cf. Section 3). Translations of the source terms of gold-test set were manually validated and classified in accordance with the set of fine-grained error categories (RE, IE, PE, ILS, TD and REM) described above, as well as a correct category (CT: correct translation). This was accomplished with the human evaluator.

The manual evaluation was carried out using the GUI that randomly displays a source sentence and its reference translation from gold-test set, 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 categories and the sole correct category described above.

The evaluator, a native Hindi speaker with excellent English skills, was instructed to follow the following criteria for evaluating the translation of a source term: (a) judge the correctness/incorrectness of the translation of the source term in the hypothesis and label it with an appropriate category listed in the GUI; (b) do not judge the whole translation, but instead, look at the local context in which both the 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 both MT system types. We measured agreement in manual classification of terminology translation. For this, we randomly selected an additional 100 segments from the gold-test set and hired another evaluator who was also a native Hindi speaker with excellent English to perform the classification task. First, we considered the binary categories (correct or incorrect term translation), i.e., we counted an agreement whenever both evaluators agreed that it was a correct (or incorrect) term translation, with agreement by chance = 1/2. Second, we considered the fine-grained categories (that included the correct and error classes), i.e., we counted an agreement whenever both evaluators agreed on the class (correct/error classes) assigned to the term translation. We considered all seven categories described above as equally likely, with agreement by chance = 1/7. As far as the agreements on classification of terminology translations with the four MT systems were concerned, we found that the kappa coefficient for the binary and sub-classes ranged from 0.97–1.0 and 0.76–0.85, respectively. It is believed that a kappa coefficient between 0.6 and 0.8 represents substantial agreement, with anything above 0.8 indicating perfect agreement. In this sense, our manual term translation classification quality could be labelled as excellent.

In Table 4, we report the manual evaluation results from the four MT tasks, where we present the distributions of the terminology translations across the seven classes we considered for the manual evaluation across the MT tasks.

Table 4. Classification of terminology translations by the human evaluator in the English-to-Hindi and Hindi-to-English PB-SMT and NMT tasks; total number of terms: 3064.

	English-to-Hindi PB-SMT	**English-to-Hindi NMT**	**Hindi-to-English PB-SMT**	**Hindi-to-English NMT**
CT	2761	2811	2668	2711
RE	15	5	18	5
IE	79	77	118	76
PE	52	47	65	73
ILS	77	44	139	90
TD	53	56	38	86
REM	27	24	18	23

7.2. Automatic Evaluation Metrics

In order to measure the performance of the proposed automatic metric, TermEval, and classification model, TermCat, we made use of three widely-used standard evaluation metrics: precision, recall and F1. We describe the definitions of precision, recall and F1 measures below. For a given class, precision was measured as the ratio of the correct predictions for the class and the number of predictions of the class, as in (2):

$$\text{Precision} = \frac{\text{Number of correct predictions for the class}}{\text{Number of predictions of the class}} \tag{2}$$

The recall was measured as the ratio of the number of the correct predictions for the class and size of the reference data for that class, as in (3):

$$\text{Recall} = \frac{\text{Number of correct predictions for the class}}{\text{Size of the reference data for the class}} \tag{3}$$

F1 is a function of precision and recall, which is calculated as the harmonic mean of precision and recall, as in (4):

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

7.3. Validating TermEval

We used TermEval to evaluate the performance of our MT systems in terminology translation on the gold-test set, which included 3064 source–target term-pairs. The TermEval scores are reported in Table 5. For clarity, we also report the total number of correct term translations by the MT systems, which is, in fact, the numerator of the right side of (1) (cf. Section 5). As can be seen from the table, the English-to-Hindi PB-SMT and NMT systems correctly translated 2610 and 2680 English terms (out of a total of 3064 terms), respectively, into Hindi, resulting in TermEval scores of 0.852 and 0.875, respectively. In the other direction (i.e., from Hindi-to-English), the PB-SMT and NMT systems correctly translated 2554 and 2540 English terms (out of a total of 3064 terms), respectively, into Hindi, resulting in TermEval scores of 0.834 and 0.829, respectively. To validate the TermEval scores, we measured the correlations between them and human judgements (cf. Table 4) on terminology translation in MT, which is presented below.

Table 5. Terminology translation accuracy in TermEval.

English-to-Hindi Task		
	Correct Translation	TermEval
PB-SMT	2610	0.852
NMT	2680	0.875
Hindi-to-English Task		
PB-SMT	2554	0.834
NMT	2540	0.829

TermEval provided the number of correct term translations by the MT systems on the test set, which are reported in Table 5. Thus, we obtained the number of incorrect term translations by the MT systems. Given the reference data (human evaluation results; cf. Table 4), we obtained the actual number of correct and incorrect term translations by the MT systems. Given the numbers from Tables 4 and 5, we created contingency tables for both the English-to-Hindi and Hindi-to-English tasks and show them in Table 6. As can be seen, there were two contingency tables for each of the MT model types: the left-hand tables are for the PB-SMT tasks, and the right-hand tables are for the NMT tasks. The first row and column of each table show the numbers that were obtained from TermEval and the reference data, respectively. The numbers (correct and incorrect term translations) from the reference data were distributed over the correct and incorrect types by TermEval, as well as the other way round. As an example, the manual evaluator labelled 2761 source terms (out of 3064 total source terms) as correct translations in the English-to-Hindi PB-SMT task, and these 2761 source terms belonged to two categories (correct: 2602, incorrect: 159), as per TermEval.

Table 6. Contingency tables.

English-to-Hindi Task					
PB-SMT			NMT		
	2610	454		2680	384
2761	2602	159	2811	2677	134
303	8	295	253	3	250
Hindi-to-English Task					
PB-SMT			NMT		
	2554	510		2540	524
2668	2554	114	2711	2540	171
396	0	396	353	0	353

Given the contingency tables in Table 6, we measured the accuracy of TermEval on the English–Hindi translation task. We report the precision, recall and F1 scores in Table 7. As can be seen, we obtained roughly similar scores on all 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 in terms of the scores in Table 7.

Table 7. TermEval's performance on terminology translation evaluation tasks.

English-to-Hindi Task		
	PB-SMT	NMT
Precision	0.997	0.999
Recall	0.942	0.953
F1	0.968	0.975
Hindi-to-English Task		
	PB-SMT	NMT
Precision	1.0	1.0
Recall	0.957	0.937
F1	0.978	0.967

7.4. TermEval: Discussion and Analysis

In this section, first we discuss the scenario in which TermEval labelled those term translations as correct, which were in fact incorrect as per the human evaluation results (false positives, cf. Table 6). Then, we discuss the reverse scenario in which TermEval labelled those term translations as incorrect, which were in fact correct as per the human evaluation results (false negatives, cf. Table 6).

7.4.1. False Positives

As can be seen from the fifth row of Table 6, there were eight and three false-positives in the English-to-Hindi PB-SMT and NMT tasks, respectively. This indicates that in each case, TermEval labelled the term translation as correct, because the corresponding reference term (or one of its LIVs) was found in the hypothesis, although the manual evaluator labelled that term translation as incorrect. We verified these cases in translations with the corresponding reference terms and their LIVs. We found that in eight out of 11 cases, the LIV lists contained incorrect inflectional variations for the reference terms. These incidents can be viewed as annotation errors, as these erroneous inflectional variations for the reference terms were included in the gold-test set at the time of its creation. For the the remaining cases, we found that the English-to-Hindi PB-SMT system made the correct lexical choice for the source terms, although the meanings of their target-equivalents in the respective translations were different from those of the source terms. This can be viewed as a *cross-lingual disambiguation* problem. As an example, one of the three source terms was "victim" (reference translation "shikaar"), and the English-to-Hindi PB-SMT system made 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 TermEval to recognise such term translation errors. We keep this topic as a subject of future work.

7.4.2. False Negatives

We see from Table 6 that the number of false negatives (e.g., 159 in the English-to-Hindi PB-SMT task) across all MT tasks was much higher than the false positives. In fact, this was responsible for the slightly worse recall scores (cf. Table 7). We point out below why TermEval failed to label such term translations as correct despite the fact that they were correct translations as per the human evaluation results.

Reordering issue: Here, first we highlight the word ordering issue in term translation. As an example, a Hindi source term "khand nyaay peeth ke nirnay" (English reference term: "division bench judgement") 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 *judgement of the division bench*". Nevertheless, TermEval implements a simple word matching module that essentially failed to capture such word ordering in the target translation. In Table 8, we report the number of instances where

TermEval failed to distinguish those term translations in the PB-SMT and NMT tasks that contained all words of the reference term (or one of its LIVs), but in an order different from the reference term. As can be seen from Table 8, these numbers were reasonably high when the target language was English. In order to capture automatically a term translation whose word order is different from that of the reference term (or one of its LIVs), we need to incorporate language-specific or lexicalized reordering rules into TermEval, which we intend to investigate in the future.

Table 8. False negatives in PB-SMT and NMT tasks when TermEval failed to distinguish reordered correct term translation.

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

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 the Hindi legal term "abhikathan". TermEval failed to label this term translation as correct due to the fact that its morphological form is different from 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 identify this term translation. We show one more example in this context. Consider the source Hindi sentence "*vivaadagrast vaseeyat* hindee mein taip kee gaee hai aur yah sthaapit nahin kiya gaya hai ki vaseeyatakarta hindee padh sakata tha" and the English reference translation "the *will in dispute* is typed in hindi and it has not been established that the testator could read hindi" from the gold-test set [Hindi is a language whose first alphabet should be capital. However, we carried out experiments with lowercased characters. This is why we show this named-entity in lowercased characters in the example]. 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 and it has not been established that the editor could read 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 from that of the reference term ("will in dispute"); and the morphological form of (part of) the translation was not identical to that of (part of) the reference term. This can be viewed as a combination of "reordering" and "inflectional" issues. TermEval failed to mark this term translation as correct. In Table 9, we report the 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 reasons. In Table 9, we see a mixed set of results, i.e., such instances are more often seen in PB-SMT when the target is Hindi and the other way round (i.e., more often seen in NMT) when the target is English.

Table 9. False negatives in the PB-SMT and NMT tasks when TermEval fails to capture a correct term translation whose morphological form is different from the reference term or one of its LIVs.

False Negative (Due to term's morphological variations)			
	PB-SMT	NMT	PB-SMT ∩ NMT
English-to-Hindi	112	87	31
Hindi-to-English	75	109	48

We recall the rule that we defined while forming the 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, the 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 replacing the reference term in the reference translation, may (syntactically) be a misfit in the context of the reference translation. This is the reason why the annotators, at the time of annotation, did not consider such inflectional variation for a reference term. As a result, TermEval, which works on top of the gold-test set, failed to capture such terminology translations. These scenarios are expected to be seen more with morphologically-rich languages like Hindi. However, in our case, we see from Table 9 that these occurred with both Hindi and English. This is a challenging problem for TermEval to address. In this context, we mention the standard MT evaluation metric METEOR [58], which, to an extent, tackles the above problems (reordering and inflectional issues) with two special modules (i.e., paraphrase and stem matching modules).

Miscellaneous reasons: In Table 10, we report the number of false negatives in the PB-SMT and NMT tasks when TermEval failed to capture a correct term translation for various reasons. There were mainly three reasons:

- 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 was Hindi. In practice, many English terms (transliterated form) are often used in Hindi text (e.g., "decree" as "dikre", "tariff orders" as "tarif ordars", "exchange control manual" as "eksachenj kantrol mainual").
- Terminology translation coreferred: The translation-equivalent of a source term was not found in the hypothesis; however, it was rightly coreferred in the target translation.
- Semantically-coherent terminology translation: The translation-equivalent of a source term was not seen in the hypothesis, but its meaning was correctly transferred into the target language. As an example, consider the source Hindi sentence "sabhee apeelakartaon ne *aparaadh sveekaar nahin* kiya aur muqadama chalaaye jaane kee maang kee" and the reference English sentence "all the appellants pleaded not guilty to the charge and claimed to be tried" from the gold-test set. In this example, the reference English sentence was the literal translation of the source Hindi sentence. Here, "aparaadh sveekaar nahin" is a Hindi term, and its English translation is "pleaded not guilty". The Hindi-to-English NMT system produced 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 that the meaning of the source term "aparaadh sveekaar nahin" was preserved in the target translation.

Table 10. False negatives in the PB-SMT and NMT tasks when TermEval fails to capture correct term translations for various reasons.

False Negative (Due to Miscellaneous Reasons)			
	PB-SMT	NMT	PB-SMT ∩ NMT
English-to-Hindi	8	4	-
Hindi-to-English	2	18	-

Missing LIVs: In a few cases, the human evaluator found that the source terms were correctly translated into the target language, but the translations were neither the reference terms nor any of its LIVs. These can be viewed as annotation mistakes since the annotators omitted adding relevant LIVs for the reference term in the gold-test set. In Table 11, we report the number of false negatives in the PB-SMT and NMT tasks when TermEval failed to capture correct term translations for this reason.

Table 11. False negatives in the PB-SMT and NMT tasks when TermEval fails to capture term translations as correct due to the absence of an appropriate LIV in gold-test set.

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

7.5. Validating TermCat

We presented the translations of the gold-test set by the MT systems to TermCat for classification. We obtained predictions for the translations of the source terms appearing in the gold-test set. The number of predictions for each of the classes (RE, IE, PE, ILS and TD) is reported in Table 12 (Rows 3 and 14). Given the reference data (human evaluation results; cf. Section 7.1), we obtained the number of correct predictions by TermCat. We report these numbers in Table 12 (Rows 4 and 15). For clarity, we also report the actual numbers corresponding to each class from Table 4 in Table 12 (Rows 5 and 16). The top and bottom halves of Table 12 represent the results for the English-to-Hindi and Hindi-to-English tasks, respectively, and the left and right sides of the table represent the PB-SMT and NMT tasks, respectively. Given this information, we evaluated TermCat's performance on the terminology translation error classification task. As usual, we used precision, recall and F1 metrics for this. Additionally, we report TermCat's overall performance in terms of macro precision, recall and F1. The macro score averages the measures for individual classes. In Table 12, we report TermCat's performance for all classes (Rows 6, 7 and 8 for the English-to-Hindi task and Rows 17, 18 and 19 for the Hindi-to-English task). The macro-averaged results over all classes are shown in Rows 9, 10 and 11 for the English-to-Hindi task and in Rows 20, 21 and 22 for the Hindi-to-English task. As can seen from Table 12, in general, we obtained high recall and low precision scores for all classes. The same trends were seen with the macro-averaged results (bold). We discuss the numbers from Table 12 below, highlighting the reasons why TermCat performed below par in some areas.

Table 12. Performance of TermCat in precision, recall and F1 metrics.

		PB-SMT					**NMT**				
		RE	**IE**	**PE**	**ILS**	**TD**	**RE**	**IE**	**PE**	**ILS**	**TD**
English-to-Hindi	Predictions	15	167	83	113	72	14	132	72	67	99
	Correct Predictions	14	69	48	67	45	5	62	42	34	53
	Reference (cf. Table 4)	19	79	52	77	53	5	77	47	44	56
	Precision	73.7	40.8	57.8	59.3	62.5	35.7	39.9	58.3	50.7	54.5
	Recall	93.3	87.3	92.3	87.1	84.9	100.0	81.2	89.4	77.3	94.6
	F1	82.3	55.6	71.1	70.6	72.0	51.3	53.1	75.1	61.2	57.6
	macroP			**58.8**					**47.8**		
	macroR			**88.9**					**88.3**		
	macroF1			**70.3**					**59.6**		
		PB-SMT					**NMT**				
		RE	**IE**	**PE**	**ILS**	**TD**	**RE**	**IE**	**PE**	**ILS**	**TD**
Hindi-to-English	Predictions	32	173	77	136	92	16	162	88	89	169
	Correct Predictions	18	101	65	112	36	5	64	71	57	82
	Reference (cf. Table 4)	18	118	62	139	38	5	76	73	90	86
	Precision	56.3	62.4	80.5	82.4	39.1	31.5	39.5	80.7	62.9	48.5
	Recall	100.0	91.5	95.4	80.6	94.7	100	84.2	78.8	62.2	95.4
	F1	72.0	74.2	87.3	80.5	55.4	47.9	53.8	79.8	62.6	64.3
	macroP			**64.1**					**52.6**		
	macroR			**92.4**					**84.1**		
	macroF1			**73.8**					**61.7**		

7.6. TermCat: Discussion and Analysis

This section discusses various aspects of term translations centering on the results of Table 12. We consider each of the error classes for discussion, starting with RE.

7.6.1. Reorder Error

As far as TermCat's performance on RE was concerned, we see from Table 12 that the F1 scores were much higher in PB-SMT than in NMT. TermCat's RE module (cf. Section 6.2) was based on a simple rule, i.e., RE (reorder error) was caught when words of a multi-word reference term were found in the hypothesis in a different order. We pointed out reordering issues in terminology translation in Section 7.4.2 with an example (reference: "division bench judgement" and MT: "judgement of the division bench"). TermCat marked the term translation shown in that example with RE, which was incorrect. We looked at the manual classification outcome (cf. Section 7.1) and investigated those terms and their translations that were tagged as RE. In other words, we looked at the distributions of term translations of the RE class over the manual classes. We found that the majority of these term translations were indeed correct translations as far as the NMT tasks were concerned. This was one of the reasons why TermCat's accuracy for RE was lower in NMT. We refer back to Table 8 where we report the number of false negatives in the PB-SMT and NMT tasks when TermEval failed to distinguish a term's correct reordered form. In fact, the numbers in Table 8 represent false positives of the RE class in four MT tasks. In order to classify such term translations as correct automatically, we should rely on language-specific rules or lexicalized reordering rules, which we intend to explore in the future.

7.6.2. Inflectional Error

As far as the IE class was concerned, TermCat performed below par and produced nearly identical scores (F1 of 53.1–55.5) in all tasks except for the Hindi-to-English PB-SMT task, for which we obtained a moderate F1 score, 74.2. We looked at the manual classification outcome (cf. Section 7.1) and investigated those terms and their translations that were tagged as IE. We found that nearly half of the term translations in all MT tasks that were automatically classified as IE were manually classified as correct translations by the manual evaluator (i.e., false positives). This significantly lowered TermCat's precision scores for IE. In Section 7.4.2, we mentioned the inflectional issue in terminology translation in a few examples ("allegation" vs. "alleged", "no witness" vs. "witnesses", and "disputed will" vs. "will in dispute"). TermCat mistakenly tagged these term translations as IE, despite them being correct translations. We refer back to Table 9, where we report the number of false negatives in the PB-SMT and NMT tasks when TermEval failed to distinguish a term's correct morphological variations. In fact, the numbers in Table 9 represent false positives of the IE class on four MT tasks. It would be a challenging task for TermCat to address this problem. We also evidenced a similar story in the case of RE above. We also see from Table 12 that the recall scores for the IE class were slightly low when the target side of the translation was Hindi. For Hindi, we used a lightweight stemmer that was limited to a small number of inflectional morphemes. This could be the reason why F1 scores for IE were slightly low when the target side of the translation was Hindi.

7.6.3. Partial Error

Now, we focus on TermCat's performance on PE. We obtained moderate and good scores for this class in the Hindi-to-English and English-to-Hindi translation tasks, respectively. We looked at the distributions of term translations of the PE class over the manual classes. We found that more than half of the false-positives of the PE class in all tasks were correct translations, which led to TermCat's lower precision for PE. We also identified a reason for this. At the time of creation of the gold-test set (cf. Section 4), human annotators omitted adding relevant variations (lexical and inflectional) for some reference terms. Most of the correct translations were tagged as PE for this reason. However, this problem can be minimised with a better termbank. In the future, we aim to make the gold-test set as exhaustive as possible by adding missing variations for the respective reference terms.

7.6.4. Incorrect Lexical Selection

Now, we discuss how TermCat performed on the ILS class. We see from Table 12 that TermCat's performance on ILS was better in PB-SMT than in NMT. In order to trace incorrect lexical selection in terminology translation, TermCat made use of external knowledge sources. TermCat's ILS module gathers translation-equivalents for the source terms using WordNet and the PB-SMT phrase table (cf. Section 6.2). We wanted to see whether the use of the PB-SMT phrase table had any impact on raising TermCat's performance in PB-SMT tasks. Hence, we counted the number of source terms that were labelled as ILS and traced using the PB-SMT phrase table. Given the manually-classified data, we divided this number into two measures: true-positives and false-positives, which are reported in Table 13. Numbers inside the square bracket (cf. Table 13) represent the total number of true-positives and false-positives of the ILS class. We see from Table 13 that the percentage of true-positives due to the use of the PB-SMT phrase table with respect to the total number of true-positives was much higher in PB-SMT than in NMT, whereas the false-positive scores were comparable in both tasks. Therefore, we can clearly see that the use of the PB-SMT phrase table in the ILS module appeared to be responsible for the difference in the classification accuracies in PB-SMT and NMT.

Table 13. True-positives and true-negatives of ILS while using the PB-SMT t-table. Numbers inside square brackets represent the total number of true-positives and false-positives of ILS.

	True-Positives		**False-Positives**	
PB-SMT	138 [177]	78.0%	50 [66]	75.7%
NMT	51 [84]	60.7%	44 [60]	73.3%

We looked at the distributions of false-positives of ILS over the manual classes. We found that they were nearly identical on all four translation tasks. In fact, most of the false-positives were correct term translations. We explained above why term translations that were labelled as PE were classified as correct term translations by the human evaluator. We pointed out that the annotators omitted adding relevant variations for a few reference terms in the gold-test set, which was the main cause of PE's low precision scores. The same story is also applicable for ILS. However, this can be minimised with an improved reference termbank. In summary, in order to improve TermCat's performance on ILS or PE, we need to make the reference termbank as exhaustive as possible.

7.6.5. Term Drop

We see from Table 12 that TermCat's performance on TD was below par. When we see the distributions of TD's true-negatives across the manual classes, we see they were mainly ILS and REM in both PB-SMT and NMT tasks. Let us recall TermCat's ILS module from Section 6.1. Any unlabelled source terms that come from the ILS module were, by default, labelled as TD (see Figure 3). Therefore, where the ILS module failed to detect an incorrect lexical selection (or partial incorrect lexical selection) for a term translation, this in turn became a true-negative of TD.

8. Conclusions

In this study, we presented a faster and less expensive semi-automatic annotation strategy following which one can create a reusable gold standard for evaluating terminology translation in MT. We also introduced an automatic evaluation metric, TermEval, for evaluating terminology translation in MT, which worked on top of the gold standard. This evaluation strategy can help MT users quickly assess terminology translation quality in automatic translation. The paper also presented a classification framework, TermCat, that can automatically classify terminology translation-related errors in MT into a set of predefined error categories. TermCat can aid the end-users in quickly identifying specific problems in relation to terminology translation in MT.

We created a gold standard test set from the English–Hindi judicial domain parallel corpus, which was used for validating TermEval and TermCat in four MT tasks. We found that TermEval represented a promising metric for the evaluation of terminology translation in MT, showing very high correlations with the human judgements in all translation tasks (i.e., F1 ranging from 96.7–97.8). We examined why TermEval failed to distinguish term translations in a few cases and identified reasons (e.g., reordering and inflectional issues in term translation) for such errors. As far as the TermCat's classification accuracy was concerned, we obtained high recall and low precision scores for all classes. We obtained reasonable accuracies with PE (F1: 71.1–87.3) in all translation tasks, with RE (F1: 72.0–82.3) and ILS (F1: 70.6–80.5) in the PB-SMT tasks, and with TD (F1 of 72.0) and IE (F1 of 74.2) in the English-to-Hindi and Hindi-to-English PB-SMT tasks, respectively, as well as moderate accuracies with IE (F1: 53.1–55.6), TD (F1: 55.4–64.3), RE (F1: 47.9–51.3) and ILS (F1: 61.2–62.6) on the respective remaining translation tasks. In sum, the classifier's performance varied across the MT models, translation directions and morphological nature of the languages involved in translation. We identified several reasons for the variations. We found that the quality of the linguistic processors and language resources and the use of external knowledge (i.e., PB-SMT phrase table) impacted TermCat's performance. As far as the classifier's performance over all classes (macro-averaged scores) was concerned, we saw that the classifier performed reasonably and competently in the terminological error classification tasks (macroF1 of 70.3 and 73.8 on the PB-SMT tasks and 59.6 and 61.7 on the NMT tasks). As expected, we saw similar characteristics across the four translation tasks with the macro-averaged recall and precision measures, i.e., high macroR and low macroP scores. We also saw that the classifier's overall accuracy (i.e., macroF1 scores) was moderately better in PB-SMT than in NMT.

TermCat had a minimal language dependency, such as a stemmer for the target language and lexical knowledge base. Stemmers are usually available for many languages; even if they are not available, they can easily be prepared like the one we used for Hindi in our work. Lexical knowledge bases are also available for a handful of languages; even if they are not available, a PB-SMT phrase table or lexicon can, to some extent, suffice. Note that any language resources that can avail of translation-equivalents for domain terms can be exploited in the classifier's lexical error identification module. Since the proposed classification model had minimal language dependency, its applicability can be extended to other languages including low resource languages.

Although our classification model, TermCat, could automatically expose specific problems in term translation in MT, it worked on top of the gold standard evaluation test set. Our term annotation (i.e., creation of the gold standard) process was not fully automated since it required a certain level of manual intervention due to many factors *vis-ˀa-vis* terminology translation, e.g., the degree of ambiguity, the morphological nature of language (i.e., related to term's inflectional variations), the data domain (i.e., related to term's lexical variations), and the quality of automatically-extracted bilingual terminology. If a fully-automated solution to this problem were to appear in the future (i.e., automatic creation of the gold standard dataset), then there would be no human dependency in the pipeline, and the proposed classification model could fully automate the nature of the terminological errors in MT.

The created gold standard test set can be regarded as an important language resource in MT research. The gold standard can also be used for the evaluation of 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 dataset, which would provide insights for such an annotation scheme.

In the future, we aim to make our gold standard evaluation test set as exhaustive as possible by amending missing lexical and inflectional variations for reference terms. We also plan to test our evaluation technique with different language pairs and domains.

Author Contributions: Investigation, R.H., M.H. and A.W.; writing, original draft, R.H. and M.H.; writing, review and editing, A.W.

Funding: 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 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.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

TW	Translation workflow
TSP	Translation service provider
MT	Machine translation
LIV	Lexical and inflectional variation
SMT	Statistical machine translation
NMT	Neural machine translation
PB-SMT	Phrase-based statistical machine translation
RE	Reorder error
IE	Inflectional error
PE	Partial error
ILS	Incorrect lexical selection
TD	Term drop
CT	Correct translation
REM	Remaining class

References

1. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the ACL-2002: 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.

2. Haque, R.; Hasanuzzaman, M.; Way, A. TermEval: An automatic metric for evaluating terminology translation in MT. In Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing, La Rochelle, France, 7–13 April 2019.
3. 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), Dublin, Ireland, 23 August 2014; pp. 42–51.
4. Haque, R.; Penkale, S.; Way, A. TermFinder: Log-likelihood comparison and phrase-based statistical machine translation models for bilingual terminology extraction. *Lang. Resour. Eval.* **2018**, *52*, 365–400. [CrossRef]
5. Devanagari. Available online: https://en.wikipedia.org/wiki/Devanagari (accessed on 28 August 2019).
6. Junczys-Dowmunt, M.; Dwojak, T.; Hoang, H. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. *arXiv* **2016**, arXiv:1610.01108.
7. Kunchukuttan, A.; Mehta, P.; Bhattacharyya, P. The IIT Bombay English—Hindi Parallel Corpus. *arXiv* **2017**, arXiv:1710.02855.
8. Koehn, P.; Och, F.J.; Marcu, D. Statistical Phrase-based Translation. In Proceedings of the HLT-NAACL 2003: Conference Combining Human Language Technology Conference Series and the North American Chapter of the Association For Computational Linguistics Conference Series, Edmonton, AB, Cananda, 27 May–1 June 2003; pp. 48–54.
9. Kalchbrenner, N.; Blunsom, P. Recurrent Continuous Translation Models. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), Seattle, WA, USA, 18–21 October 2013; pp. 1700–1709.
10. 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), Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
11. 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, Montreal, QC, Cananda, 8–13 December 2014; pp. 3104–3112.
12. 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, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
13. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
14. Hassan, H.; Aue, A.; Chen, C.; Chowdhary, V.; Clark, J.; Federmann, C.; Huang, X.; Junczys-Dowmunt, M.; Lewis, W.; Li, M.; et al. Achieving Human Parity on Automatic Chinese to English News Translation. *arXiv* **2018**, arXiv:1803.05567.
15. Sag, I.A.; Baldwin, T.; Bond, F.; Copestake, A.; Flickinger, D. Multiword expressions: A pain in the neck for NLP. In Proceedings of the CICLing 2002, the Third International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science, Mexico City, Mexico, 17–23 February 2002; Gelbukh A., Ed.; Springer-Verlag: Berlin/Heidelberg, Germany, 2002.
16. Mitkov, R.; Monti, J.; Pastor, G.C.; Seretan, V. (Eds.) *Multiword Units in Machine Translation and Translation Technology, Current Issues in Linguistic Theory*; John Benjamin Publishing Company: Amsterdam, The Netherlands, 2018; Volume 341.
17. Haque, R.; Hasanuzzaman M.; Way A. Multiword Units in Machine Translation—Book Review. *Mach. Transl.* **2019**, *34*, 1–6. [CrossRef]
18. Rigouts Terryn, A.; Hoste, V.; Lefever, E. In no uncertain terms: A dataset for monolingual and multilingual automatic term extraction from comparable corpora. *Lang. Resour. Eval.* **2019**. [CrossRef]
19. 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), Madrid, Spain, 19–22 June 2012; pp. 193–208.
20. 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), Vancouver, BC, USA, 22–26 October 2014; pp. 54–68.

21. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, 23–25 May 2012; pp. 2214–2218.
22. BitterCorpus. Available online: https://hlt-mt.fbk.eu/technologies/bittercorpus (accessed on 28 August 2019).
23. Pazienza, M.T.; Pennacchiotti, M.; Zanzotto, F.M. Terminology extraction: An analysis of linguistic and statistical approaches. In *Knowledge Mining*; Sirmakessis, S., Ed.; Springer: Berlin/Heidelberg, Germany, 2005; pp. 255–279.
24. 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, Alacant/Alicante, Spain, 28–30 May 2018; pp. 149–158.
25. Popović, M.; Ney, H. Towards Automatic Error Analysis of Machine Translation Output. *Comput. Linguist.* **2011**, *37*, 657–688. [CrossRef]
26. Bentivogli, L.; Bisazza, A.; Cettolo, M.; Federico, M. Neural versus Phrase-Based Machine Translation Quality: A Case Study. *arXiv* **2016**, arXiv:1608.04631.
27. 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. *Prague Bull. Math. Linguist.* **2017**, *108*, 159–170. [CrossRef]
28. Macketanz, V.; Avramidis, E.; Burchardt, A.; Helcl, J.; Srivastava, A. Machine Translation: Phrase-based, Rule-Based and Neural Approaches with Linguistic Evaluation. *Cybern. Inf. Technol.* **2017**, *17*, 28–43. [CrossRef]
29. 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 the MT Summit XVI: The 16th Machine Translation Summit, Nagoya, Japan, 18–22 September 2017; pp. 55–71.
30. Lommel, A.R.; Uszkoreit, H.; Burchardt, A. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics. *Tradumática* **2014**, *12*, 455–463. [CrossRef]
31. 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 the EAMT User Studies and Project/Product Descriptions, Prague, Czech Republic, 29–31 May 2017; pp. 41–46.
32. Vintar, Š. Terminology Translation Accuracy in Statistical versus Neural MT: An Evaluation for the English–Slovene Language Pair. In 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, Vancouver, BC, Canada, 2–7 October 2018; Du, J., Arčan, M., Liu, Q., Isahara, H., Eds.; European Language Resources Association (ELRA): Miyazaki, Japan, 2018; pp. 34–37.
33. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv* **2016**, arXiv:1609.08144.
34. Karstology. Avaialbe online: https://en.wiktionary.org/wiki/karstology (accessed on 28 August 2019).
35. 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 the RANLP 2019: Recent Advances in Natural Language Processing, Varna, Bulgaria, 2–4 September 2019; (to appear).
36. Huang, G.; Zhang, J.; Zhou, Y.; Zong, C. A Simple, Straightforward and Effective Model for Joint Bilingual Terms Detection and Word Alignment in SMT. In *Natural Language Understanding and Intelligent Applications, ICCPOL/NLPCC 2016*; Springer: Cham, Switzerland, 2016; Volume 10102, pp. 103–115.
37. Koehn, P.; Hoang, H.; Birch, A.; Callison-Burch, C.; Federico, M.; Bertoldi, N.; Cowan, B.; Shen, W.; Moran, C.; Zens, R.; et al. Moses: Open Source Toolkit for Statistical Machine Translation. In Proceedings of the Interactive Poster and Demonstration Sessions, Prague, Czech Republic, 25–27 June 2007; pp. 177–180.
38. James, F. Modified Kneser-Ney Smoothing of N-Gram Models. Available online: https://core.ac.uk/download/pdf/22877567.pdf (accessed on 28 August 2019).
39. Heafield, K.; Pouzyrevs.ky, 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), Sofia, Bulgaria, 4–9 August 2013; pp. 690–696.

40. 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, Seattle, WA, USA, 18–21 October 2013; pp. 1387–1392.
41. 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, Portland, OR, USA, 19–24 June 2011; pp. 1045–1054.
42. Och, F.J.; Ney, H. A Systematic Comparison of Various Statistical Alignment Models. *Comput. Linguist.* **2003**, *29*, 19–51. [CrossRef]
43. 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, Montréal, QC, Canada, 3–8 June 2012; pp. 427–436.
44. 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, Prague, Czech Republic, 23–30 June 2007; pp. 144–151.
45. Junczys-Dowmunt, M.; Grundkiewicz, R.; Dwojak, T.; Hoang, H.; Heafield, K.; Neckermann, T.; Seide, F.; Germann, U.; Fikri Aji, A.; Bogoychev, N.; et al. Marian: Fast Neural Machine Translation in C++. In Proceedings of the ACL 2018, System Demonstrations; Association for Computational Linguistics, Melbourne, Australia, 26–31 July 2018; pp. 116–121.
46. Gage, P. A New Algorithm for Data Compression. *C Users J.* **1994**, *12*, 23–38.
47. 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), Berlin, Germany, 7–12 August 2016; pp. 1715–1725.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
49. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
50. Press, O.; Wolf, L. Using the Output Embedding to Improve Language Models. *arXiv* **2016**, arXiv:1608.05859.
51. Gal, Y.; Ghahramani, Z. A Theoretically Grounded Application of Dropout in Recurrent Neural Networks. *arXiv* **2016**, arXiv:1512.05287.
52. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
53. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. *arXiv* **2015**, arXiv:1511.06709.
54. Poncelas, A.; Shterionov, D.; Way, A.; de Buy Wenniger, G.M.; Passban, P. Investigating Backtranslation in Neural Machine Translation. In Proceedings of the 21st Annual Conference of the European Association for Machine Translation (EAMT 2018), Alacant/Alicante, Spain, 28–30 May 2018; pp. 249–258.
55. 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 Language Resources and Evaluation Conference (LREC'14)*, Reykjavik, Iceland, 26–31 May 2014; pp. 3550–3555.
56. Koehn, P. Europarl: A parallel corpus for statistical machine translation. In Proceedings of the MT Summit X: The Tenth Machine Translation Summit, Phuket, Thailand, 12–16 September 2005; pp. 79–86.
57. Moses Tokeniser. Available online: https://github.com/moses-smt/mosesdecoder/blob/master/scripts/tokenizer/tokenizer.perl (accessed on 28 August 2019).
58. 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, Edinburgh, UK, 30–31 July 2011; pp. 85–91.
59. Snover, M.; Dorr, B.; Schwartz, R.; Micciulla, L.; Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA-2006), Cambridge, MA, USA, 8–12 August 2006; pp. 223–231.
60. Koehn, P. Statistical Significance Tests for Machine Translation Evaluation. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), Barcelona, Spain, 25–26 July 2004; Lin, D., Wu, D., Eds.; 2004; pp. 388–395.
61. 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), Leuven, Belgium, 30–31 May 2011; pp. 35–40.

62. SDL Trados Studio. Available online: https://en.wikipedia.org/wiki/SDL_Trados_Studio (accessed on 28 August 2019).
63. PyQt. Available online: https://en.wikipedia.org/wiki/PyQt (accessed on 28 August 2019).
64. Gold Standard Data Set (English–Hindi). Available online: https://www.computing.dcu.ie/~rhaque/termdata/terminology-testset.zip (accessed on 28 August 2019).
65. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Measur.* **1960**, *20*, 37–46. [CrossRef]
66. Porter, M.F. An algorithm for suffix stripping. *Program* **1980**, *14*, 130–137. [CrossRef]
67. Ramanathan, A.; Rao, D. Lightweight Stemmer for Hindi. In Proceedings of the EACL 2003 Workshop on Computational Linguistics for South-Asian Languages—Expanding Synergies with Europe, Budapest, Hungary, 12–17 April 2003; pp. 42–48.
68. Fellbaum, C. *WordNet: An Electronic Lexical Database*; Language, Speech, and Communication; MIT Press: Cambridge, MA, USA, 1998.
69. Narayan, D.; Chakrabarti, D.; Pande, P.; Bhattacharyya, P. An Experience in Building the Indo WordNet—A WordNet for Hindi. In Proceedings of the First International Conference on Global WordNet (GWC2002), Mysore, India, 21–25 January 2002; p. 8.