Revisit Automatic Error Detection for Wrong and Missing Translation – A Supervised Approach

Wenqiang Lei[1, 2], **Weiwen Xu**[2], **Ai Ti Aw**[2], **Yuanxin Xiang**[2], **Tat-Seng Chua**[1]
[1]National University of Singapore, Singapore
[2]Institute for Infocomm Research, A*STAR, Singapore
wenqianglei@gmail.com
{xuwl, aaiti, xiang_yuanxin}@i2r.a-star.edu.sg
dcscts@nus.edu.sg

Abstract

While achieving great fluency, current machine translation (MT) techniques are bottle-necked by adequacy issues. To have a closer study of these issues and accelerate model development, we propose automatic detecting adequacy errors in MT hypothesis for MT model evaluation. To do that, we annotate missing and wrong translations, the two most prevalent issues for current neural machine translation model, in 15000 Chinese-English translation pairs. We build a supervised alignment model for translation error detection (*AlignDet*) based on a simple *Alignment Triangle* strategy to set the benchmark for automatic error detection task. We also discuss the difficulties of this task and the benefits of this task for existing evaluation metrics.

1 Introduction

Different translation errors impact translation comprehensibility and adequacy differently. For example, wrong product or terminology translation can be considered more severe than missing translation in the eCommerce domain. Though many machine translation evaluation (MTE) metrics have been proposed, most of them are not able to provide a direct connection between the score and the error class to emphasize the different impact on translation comprehension and adequacy. Most MTE scores measure the similarity between the MT hypothesis and the reference using n-gram and are useful in providing immediate feedback on model performance for system development and objective evaluation for system comparison. To further analyse the MT hypothesis, fine-grained translation analysis is normally conducted manually which involves expensive human effort (Popović and Ney, 2011; Popović, 2011b; Fishel et al., 2012; Vilar et al., 2006).

This motivates early unsupervised error detection methods (Popović and Ney, 2011). Such methods find the erroneous words by calculating the edit distance between the MT hypothesis and its corresponding reference. However, such unsupervised methods suffer from low accuracy and inability to provide and distinguish certain error classes like *Missing words* and *Wrong translation*[1] (c.f. Sec 3). On the other hand, the works on supervised methods are few due to limited training resources available. The annotated corpus are mainly collected from outputs of Phrase-based Machine Translation systems and usually suffer from small quantity (around 200 instances per language pair) (Fishel et al., 2012), noisy contents (derived from student assignments) (Wisniewski et al., 2014) or overabundant efforts (additional human efforts are required on postediting the MT hypothesis) (Popović and Arcan, 2016), making the progress and development of this task challenging.

We believe a corpus annotated with different translation error classes will facilitate the research on translation error detection. Therefore, we construct a high quality annotated corpus (*TransErr*) comprising 15000 Chinese-English translation pairs with inter-annotator agreement at 0.804 measured by Cohen's Kappa (Cohen, 1960). Different from existing error detection works which focus on all error classes, we currently only take care of missing and wrong translation (Specia et al., 2018), the major errors related to adequacy, which is a wide-known issue in neural machine translation (NMT) (Zheng et al., 2019). The errors tags are annotated on source (Chinese) sentences to reflect the loyalty and adequacy with respect to the source. Based on *TransErr*, we discuss the error distribution of dif-

[1] The *wrong translation* is referred to as *incorrect lexical choice* in (Popović and Ney, 2011)

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 942–952,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

ferent systems and challenges of unsupervised error annotations using post-edits.

TransErr enables the study task of supervised MT error detection. We benchmark this task by providing a simple yet efficient model called (*AlignDet*). *AlignDet* is based on our observation of the *Alignment Triangle* whose basic idea is that if a word is well translated, the translated word and the corresponding word in reference should be equivalent. We implement such observation by employing both monolingual and bilingual alignment systems as well as various features. We further provide error analyses of *AlignDet* to give more insight to this task.

To study the application of this task, we further conduct a discussion on the impact of different error classes on MT evaluation metrics based on WMT Chinese–English Direct Assessment (DA) corpus (Bojar et al., 2017). We find that embracing the results of our simple *AlignDet* model in existing MTE metrics helps to achieve significantly better correlation with human. Our further analysis on gold standard error annotation suggests that wrong translation tends to produce worse translation. This discovery gives further credence for the usefulness of our proposed adequacy-oriented error detection task.

In short, the paper is towards carrying out the first systematic study of supervised approaches for error detection, with the goal of accelerating the research and development of MT towards human translation quality. Our systematic contribution lies on annotated data, supervised model and discussions of its potential to help existing machine translation evaluation task. Specifically:

- We propose a task of adequacy-oriented error detection for machine translation evaluation and contribute a high-quality corpus *TransErr* annotated by professional annotators for machine translation. The corpus are available[2].
- We propose an *AlignDet* to set the benchmark of this task based on our key observation of *Alignment Triangle*. As far as we know, this is the first model making use of all three alignments from source, MT hypothesis and reference concurrently.
- We conduct various discussions on the challenges of this task and its potential to help existing machine translation evaluation task.

2 Related Work

Existing automatic MTE focus on measuring the similarity between MT hypothesis and references. Various dimensions have been considered, such as lexical level (Papineni et al., 2002; Snover et al., 2009; Lo, 2017; Popović, 2017), syntax level (Owczarzak et al., 2007; Duma and Menzel, 2017; Liu and Gildea, 2005), semantic level (Stanojević and Sima'an, 2015; Shimanaka et al., 2018) and discourse level (Guzmán et al., 2014). However, such methods give an overall score to a system or give a relative ranking to a pair of systems, without any detailed insight of MT system output.

To get the fine-grained insights of the MT model performance, Popović and Ney (2011); Popović et al. (2011a) make the first step towards automatic translation error detection using unsupervised approach by performing monolingual alignment between a MT hypothesis and its corresponding reference through WER (Levenshtein, 1966) and PER (Tillmann et al., 1997). Popović (2011b) further develop this algorithm into an open-source tool and demonstrates that the detected errors helps to build better evaluation metrics. However, the edit distance algorithm is not robust (*c. f.* Sec. 3.2). Zeman et al. (2011) improve the monolingual alignment by borrowing ideas from HMM (Vogel et al., 1996) but the error detection performance is still far from real application. To facilitate this task, Wisniewski et al. (2014) derive corpus of 4,854 source sentences from the assignment of master student specializing on translation but the data turns out to be too noisy.

At the same time, edit distance is carried over to another shared task in WMT (Callison-Burch et al., 2012; Specia et al., 2018; Fan et al., 2018; Vaswani et al., 2017) called word-level Quality Estimation (QE) (Luong et al., 2013; Han et al., 2013; Wisniewski et al., 2014; Kim et al., 2017). In QE tasks, the edit distance is calculated between a MT hypothesis and its post-edits. Word needs to be edited is tagged as BAD while the unedited words are tagged as GOOD. The task requires participants to predict the label for each word. Although similar to our settings, our error detection

[2] Please contact us to get the *TranErr* corpus and the *AlignDet* source code. Before that, you need to have access to LDC NIST MT evaluation sets {LDC2005T06, NIST02, NIST03, NIST04, NIST05, NIST08}.

task differs from the word-level QE task in the following aspects:

- The QE task is for confidence estimation (Specia et al., 2010) which estimates whether an output is good for an end user. Our error detection task is for fine-grained evaluation and comparison during model development.
- QE only have `GOOD` and `BAD` tag on the MT hypothesis. It cannot make robust distinction on different error classes due to the intrinsic shortcomings of the edit distance algorithm.
- Existing QE data does not have reliable label on the source sentence to indicate which source word cause the errors (*c.f.* Sec. 3.2), which is claimed as "particularly important to translation adequacy" in WMT 18 (Specia et al., 2018) and WMT 19[3].

3 TranErr Corpus and Analysis

TransErr corpus comprises 15000 triples <source, MT hypothesis, reference> with source sentences extracted from LDC NIST MT evaluation sets {LDC2005T06, NIST02, NIST03, NIST04, NIST05, NIST08}. MT hypothesis is obtained through an NMT engine trained on 25 million Chinese-English parallel sentences. To study the correlation between error detection and human evaluation, we also conduct additional annotation on WMT17 Direct Assessment (DA) dataset containing 560 Chinese-English translation pairs.

Annotation is conducted by two professional translators who specialize on Chinese-English translation. All three information <source, MT hypothesis, reference> are given during annotation. Annotators are instructed to mark on the source the two adequacy related errors (*Missing & Wrong*) found in the MT hypothesis and to follow the minimum span principle, i.e. identifying the minimum number of words whose content are not (correctly) conveyed in the MT hypothesis regardless of fluency and grammatical correctness. If a span in the source is incorrectly translated in the MT hypothesis, all words in that span will be marked as *Wrong* (`W`); if totally missing, it will be marked as *Missing* (`M`). Due to the intrinsic difference between the two languages, not all words in the source need to be translated in the MT hypothesis. In this case, we do NOT annotate them as missing. We also have more stringent conditions for proper nouns and terminologies since their translations are more specific and domain dependent. They have to be unerringly correct or otherwise, marked as *Wrong Terminology* (`WT`) or *Missing Terminology* (`MT`[4]). All correctly translated words are automatically assigned a label `OK`.

Annotators are first asked to annotate 100 sentence pairs followed by a discussion among the project team to resolve any disagreement. After 5 such iterations, we start the formal annotation. 10% overlapped instances are given to each annotator for monitoring the inter-annotation agreement (IAA). The disagreed instances are discussed every day. We obtain an overall IAA of 0.804 measured by Cohen's Kappa. The speed for both annotators is about 50 sentence pairs per hour.

3.1 Translation Error Analysis

Table 1 shows the error distribution of *TransErr* and WMT 17 corpus. The corpora have different BLEU with *TransErr* having more missing (`M`) errors and WMT17 more wrong (`W`) errors. As WMT17 contains output from multiple MT systems, we further investigate the errors distributions among the different systems, in particular the output from AFRL and NRC. AFRL and NRC have comparable BLEU but different human evaluation. Error distributions for the two corpora vary quite significantly with AFRL having more `W` errors (45% `W` and 19% `WT`) while NRC has more `M` errors (31% `M` and 31% `MT`). This demonstrates the importance of classifying translation errors and the need for the errors to be considered during evaluation.

3.2 Challenges of Automatic Translation Error Annotation

Though MTE and error detection are two different tasks, they are closely related. Popović and Ney (2011) presented the first framework for automatic analysis and classification of translation errors using unsupervised approach for SMT systems. The work has been evolved to word-level QE task in WMT. Instead of directly annotating errors on the source as in our proposed task, we re-implement

[3] http://www.statmt.org/wmt19/

[4] In this paper, we conventionally use the normal font MT to represent *machine translation* while use the typewriter font `MT` to stand for *Missing Terminology*.

	Human	BLEU	Sen	Word	OK	W	WT	M	MT
TransErr	-	19.1	15000	403691	388372	4529 / 0.29	1694 / 0.11	8234 / 0.54	862 / 0.06
WMT17	-	16.2	560	14090	13025	410 / 0.39	206 / 0.19	320 / 0.30	129 / 0.12
AFRL	-0.016	20.6	34	917	820	44 / 0.45	19 / 0.19	17 / 0.18	17 / 0.18
NRC	0.079	20.3	32	786	698	17 / 0.19	17 / 0.19	27 / 0.31	27 / 0.31

Table 1: Corpus Statistics for *TransErr*, WMT17, AFRL and NRC. AFRL and NRC are subsets of WMT17. Human is the corresponding system-level DA score assigned by human in WMT17. All BLEU are calculated using **one reference**. Sen and Word columns list the number of sentences and words while the other four columns present the number of each error class and its proportion.

Label	OK	W	WT	M	MT
F1	0.93	0.33	0.30	0.43	0.35

Table 2: F1 for each class annotated by the unsupervised method using our manual annotation as gold reference

QE and follow (Popović and Ney, 2011) to derive source annotation through two stages of alignments. The first stage is monolingual alignment between post-edits and MT hypotheses through calculating their edit distance to find missing and wrong translations. The second stage is bilingual alignment using Giza++ (Och and Ney, 2000) on post-edits and the sources to propagate the error labels on the post-edits to the corresponding sources. Similar to our manual annotation, we assign MT and WT label for terminologies.

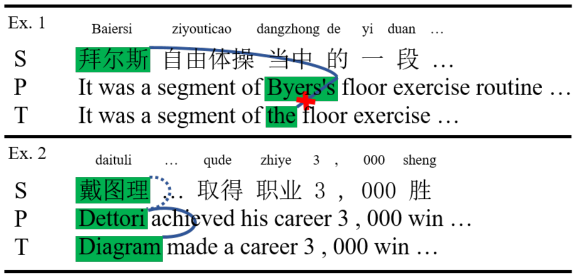
Ex. 1
Baiersi ziyouticao dangzhong de yi duan …
S 拜尔斯 自由体操 当中 的 一 段 …
P It was a segment of Byers's floor exercise routine …
T It was a segment of the floor exercise …
Ex. 2
daituli … qude zhiye 3 , 000 sheng
S 戴图理 … 取得 职业 3 , 000 胜
P Dettori achieved his career 3 , 000 win …
T Diagram made a career 3 , 000 win …

Figure 1: Examples of unsupervised annotation on WMT17. (S), (P) and (T) refer to source, post-edits and MT hypothesis respectively. Solid line indicates correct alignment. Solid line with a red cross refers to wrong alignment while dotted line refers to missing alignment. This convention applies to the whole paper.

To get empirical study, we perform unsupervised annotation on our WMT17 dataset using the above-mentioned method. Table 2 demonstrates its F1 score using the manual annotation as the gold standard. We investigate and summarize the causes in Figure 1:

- The edit distance algorithm is not robust in error class detection since it is purely based on symbolic comparison. (Ex. 1) illustrates a case where *Byers's* is missing. However, TER tool tags it as a substitution of *the* in the MT hypothesis because there is a perfect symbolic alignment before and after the erroneous word. This intrinsic inability to detect missing and wrong translation is also discussed in (Popović and Ney, 2011).

- Giza++ cannot generate reliable alignments for low-frequency words (Riley and Gildea, 2010) and such low-frequency words contribute to a large extent of erroneous words. These misalignments propagate from the post-edits to the source sentence and lead to incorrect error tags. (Ex. 2) gives an example, the *Dettori* should be aligned to *daituli* but was not aligned to any word by GIZA++. This leads to an incorrect OK tag in the source sentence.

As such, unsupervised error annotation presents many challenges even leveraging on post-edits. Hence a set of more reliable annotated data will be able to contribute much to the error detection task.

4 AlignDet Model

To revive fine-grained translation error detection, we propose *AlignDet*, a supervised feature-based solution using *TransErr* to set the benchmark for this task. We discuss our *Alignment Triangle* observation and briefly list the features we used. In the subsequent discussions, we denote s_i, t_j and r_k as the ith, jth and kth word in a source sequence S, MT hypothesis T and reference R separately.

4.1 Alignment Triangle Observation

In translation, adequacy can be simply defined as all contents in the source are reproduced correctly

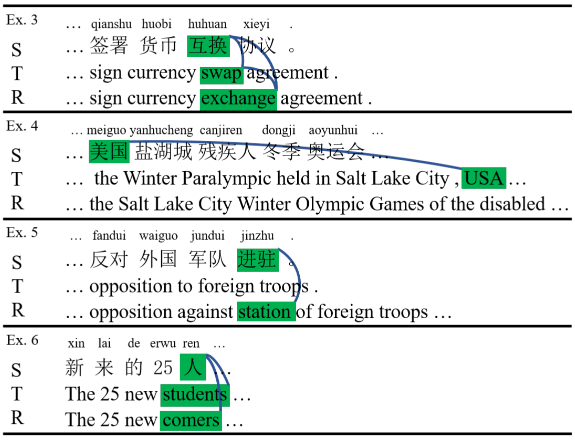
Ex. 3
… qianshu huobi huhuan xieyi .
S … 签署 货币 互换 协议 。
T … sign currency swap agreement .
R … sign currency exchange agreement .
Ex. 4
… meiguo yanhucheng canjiren dongji aoyunhui …
S … 美国 盐湖城 残疾人 冬季 奥运会 …
T … the Winter Paralympic held in Salt Lake City , USA …
R … the Salt Lake City Winter Olympic Games of the disabled …
Ex. 5
… fandui waiguo jundui jinzhu .
S … 反对 外国 军队 进驻 。
T … opposition to foreign troops .
R … opposition against station of foreign troops …
Ex. 6
xin lai de erwu ren …
S 新 来 的 25 人 …
T The 25 new students …
R The 25 new comers …

Figure 2: Four samples from *TransErr*. For each example, we show the source (S), MT hypothesis (T), reference (R).

in the MT hypothesis. As long as the MT hypothesis and the reference are semantically equivalent, such reproduction may not require the MT hypothesis and the reference to be exactly identical. This motivates us to analyze the translation t_j in T and r_k in R for the same s_i in S[5], in the following error detection scenarios. We refer t_j and r_k for the same s_i as translation correspondence.

Pattern1: s_i has translation correspondence t_j and r_k. It should be labeled as `OK` if t_j and r_k are semantically equivalent; otherwise, it should be labeled as `W`. For example, in (Ex. 3) from Figure 2, the two words, *swap* and *exchange*, are equivalent in the given context, which should be marked as `OK`. However, compared to (Ex. 6), *students* and *comers* are not equivalent in the given context, although they may be equivalent in higher abstraction, hence it is annotated as `W`.

Pattern2: s_i does not have translation correspondence r_k. Assuming *R* does not have any adequacy issue, this would imply that s_i does not need to be explicitly translated. In this case, the omission of t_j is an `OK` translation. However, if it is explicitly translated in T as t_j, we will still accept it as an `OK` translation though it may cause minor redundancy issue and we assume that it will not harm adequacy. For example, either explicitly translation *meiguo* to *USA* or not in (Ex. 4) should be fine.

Pattern3: s_i has translation correspondence r_k but not t_j. This implies that translation of s_i is missing and should be translated. A `M` label will be given in this case. (Ex. 5) shows an example of this. *jinzhu* does not have t_j but it is necessary for it to be translated as the translation correspondence is presence in *R* as r_k.

The same rationale also applies to terminologies, except that they are marked with `MT` or `WT` and have to be unerringly translated for them to be marked as `OK`. For the other cases not illustrated above, we assume that they are either not happening or do not affect the adequacy of the translation.

4.2 Features and Model

To capture the above observation, we build two sets of features, one for finding the corresponding translation t_j and r_k for each s_i, the other for checking the equivalence of t_j and r_k. Here, we only briefly list the features adopted in our model as they are all commonly used standard features. We give detailed description of all features in Apendix A.1.

Giza++ (Och and Ney, 2000) is used to find the corresponding translation t_j and r_k for each s_i. However, as Giza++ generates noisy alignments especially for low-frequency words, we propose a set of features to complement Giza++ results. The same features are applied on T and R, hence we only describe the feature for the alignment between S and T. Suppose that s_i is aligned to t_j[6]. These features are: s_i's POS tag; a binary feature of whether NER tags of s_i and t_j are the same; the corpus frequency and Giza++ translation probability of s_i and t_j; the similarity between t_j and s_i's most frequent translation; the number of aligned-pairs among words linking to s_i and t_j in dependency tree.

We leverage on the state-of-the-art monolingual alignment tool Sultan et al. (2014) to check the equivalence of t_j and r_k. The tool leverages on paraphrase lexicon (Pavlick et al., 2015) and dependency relations to find equivalent expression between two sentences in the same language. The following features are used: a binary feature indicating whether t_j and r_k are aligned by tool proposed by (Sultan et al., 2014); A binary feature indicating whether the NER tag of t_j and r_k are the same.

While this error detection problem can be modeled as sequential labelling task, we choose simple classification model as the benchmark. Therefore,

[5] Generally, t_j, r_k and s_i can be multi-word expression. For the ease of explanation for our feature extraction method, we describe each of them as a single word.

[6] Following Hu et al. (2018), we make adjustment by penalizing the frequency of both words.

we use an Mutilayer Perceptron (MLP) as the detection model which takes all above features in for each source word s_i to choose one of the labels in {`OK, W, WT, M, MT`} as the prediction.

5 Experiments

We conduct empirical experiments to evaluate the effectiveness of our error detection model, perform error analysis and discuss the challenges of this task. We split our *TransErr* dataset into training and developing sets in a ratio of 9:1 and adopt the whole WMT DA dataset as the testing set. Error detection results are reported in single-class F1. All Chinese sentences are tokenized by THULAC (Li and Sun, 2009).The rest of the pre-processings like POS, NER, parsing for both English and Chinese are performed by Stanford Corenlp (Manning et al., 2014).

5.1 Baselines

We build a naive baseline, random model, to test the worst performance. It randomly assigns a label to each source word according to the distribution of each class in the training set. We then evaluate the performance of our proposed MLP model (*AlignDet*) and two recent models under WMT word-level QE shared-task:

· **CEQE** (Hu et al., 2018) Using a three-part neural network model, which encodes semantic factors, local context and global context successively.

· **SHEF-bRNN** (Ive et al., 2018) Adopting bi-directional recurrent neural network to learn a representation for <source, MT hypothesis> sentence pairs.

We replicate these two models utilizing original implementations but modify them to work in our 5-label setting.[7] We also analyze our proposed model and its variant by ablating or substituting its components

· **Raw alignment** Using only the raw results from the monolingual and bilingual alignment systems and NER on the source to examine the effect of alignment on the model.

· **No reference** We remove all features derived from the reference, for example, bilingual alignment features between S and R, and monolingual alignment features between T and R. This is to test the efficacy of using only source information for translation performance evaluation.

5.2 Error Detection Results

Table 3 shows our error detection results on the developing set and test set. Our random baseline (Row (1)) provides a preliminary insight for this error detection problem, with extremely low F1 scores (nearly 0) except for the `OK` class. It is because errors are scarcely distributed (see Table 1). In addition, the QE models (Row 2, 3) do not perform well in error classification. This is reasonable because they are not designed for error detection in our setting. The overall better performance of our *AlignDet* model supports that our *Alignment Triangle* Observation and its contribution to the error detection task.

However, with raw alignment (Row 4 vs Row 5), the performance of *AlignDet* drops sharply. It is because the alignment model, especially Giza++, generates noisy results. This demonstrates the efficacy of other features in rectifying Giza++ results. The comparison between Row (4) and Row (6) also demonstrates the importance of reference, especially for detecting `W` label. It is because wrong translation is also likely to get aligned with the source as it can be the translation of the source words in some other contexts. However, reference gives us a pointer to the correct word choice in the given context, hence helping determining `W` label. Without reference, we may need to rely on more contextual modeling effort to detect a wrong translation.

5.3 Error Analysis

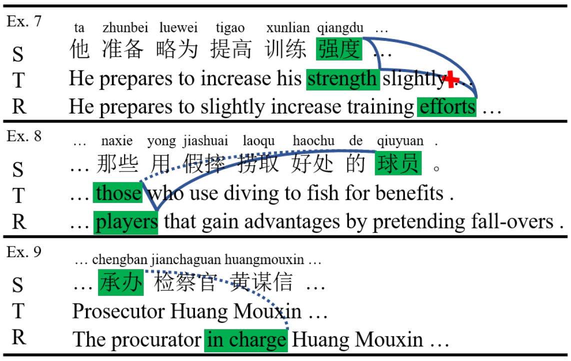
Ex. 7
ta zhunbei luewei tigao xunlian qiangdu …
S 他 准备 略为 提高 训练 强度 …
T He prepares to increase his strength slightly …
R He prepares to slightly increase training efforts …
Ex. 8
… naxie yong jiashuai laoqu haochu de qiuyuan .
S … 那些 用 假摔 捞取 好处 的 球员 。
T … those who use diving to fish for benefits .
R … players that gain advantages by pretending fall-overs .
Ex. 9
… chengban jianchaguan huangmouxin …
S … 承办 检察官 黄谋信 …
T Prosecutor Huang Mouxin …
R The procurator in charge Huang Mouxin …

Figure 3: Representative examples of the confusion cases for *AlignDet*.

To get more insight into the difficulty of this task, we perform error analysis on *AlignDet*. We only examine the confusion matrix on test set due to the limitation of space in Table 4. It shows that the significant problems are: (a) Predict `W` as `OK`

[7] We also tried QEBrain (Fan et al., 2018; Wang et al., 2018) using the released code. However, it did not converge to any meaningful predictions.

	Developing					Testing				
	OK	W	WT	M	MT	OK	W	WT	M	MT
(1) Random baseline	0.959	0.009	0.013	0.020	0.027	0.947	0.025	0.029	0.041	0.017
(2) CEQE	0.944	0.134	0.180	0.245	0.161	0.914	0.161	0.147	0.206	0.094
(3) SHEF-bRNN	0.916	0.140	0.212	0.127	0.262	0.884	0.153	0.123	0.098	0.114
(4) AlignDet	**0.973**	**0.275**	**0.480**	**0.416**	**0.365**	**0.964**	**0.212**	**0.293**	**0.400**	**0.429**
(5) Raw alignment	0.762	0.051	0.049	0.168	0.114	0.781	0.131	0.131	0.263	0.328
(6) No reference	0.977	0.040	0.448	0.289	0.289	0.962	0.075	0.213	0.360	0.378

Table 3: F1 score for each class produced by different error detection models.

Pred. Gold	OK	W	WT	M	MT
OK	12635	203	68	106	13
W	279	83	3	40	5
WT	109	28	54	9	6
M	121	54	12	122	11
MT	26	12	31	16	44

Table 4: Confusion matrix on test set.

(b) Predict OK as M (c) Predict M as OK. We illustrate the wrong prediction in Figure 3.

(Ex. 7) demonstrates a case for "W as OK". Both *strength* (the wrong translation in T) and *efforts* (the correct translation in R) are aligned to the source word *qiangdu* by Giza++. Though both of them are possible translations for *qiangdu*, they are not semantically equivalent in this context. Unfortunately, monolingual alignment tool aligns them as a pair, hence a more contextual and semantically-sensitive monolingual alignment is required. (Ex. 8) demonstrates a case for "OK as M". In T, *qiuyuan* is translated to *those* and is not detected by our bilingual alignment. This leads the model to treat it as not been translated. This example demonstrates the constraint of bilingual alignment in finding the translation correspondence, which would require contextual-level understanding. (Ex. 9) demonstrates a case for "M as OK". Similar to (Ex. 8), the mis-prediction is caused by the failure of bilingual alignment (between *chengban* in S and *in charge* in R). Our model regards *chengban* as omittable as bilingual alignment fails to align it to *in charge* in R. As a result, though *chengban* is translated in R but not translated in T, we classify it as OK instead of M. This example again demonstrates the important of alignment in error detection task.

Our qualitative error analyses highlight the bottleneck of *AlignDet* model on the proposed alignment features. In a broader context, both bilingual alignment and monolingual alignment are still having open issues in the realm of NLP (Sultan et al., 2014). An intuitive solution is turning to soft alignment (e.g., attention) using neural techniques (Liu and Sun, 2015; Tamura et al., 2014). This raises interesting and challenging questions on how to leverage on such approaches to capture our observations in finding corresponding translations t_j and r_k for each s_i and comparing their equivalence? Neural models usually require large data and how to leverage on limited error annotation data (say, 15000 sentence pairs) and large parallel corpus for such purpose?

6 Discussion: How Does Error Class Affect Existing MT Evaluation Metrics?

This paper proposes an adequacy translation error detection model with the aim to study its contribution to MT. We presume that fine grained MT evaluation is useful for assessing different properties of MT models. Hence, we go one step further to discuss how fine-grained error detection benefits existing MT evaluation metrics using DA as the base for our evaluation.

To explore further, we define a Weight Adequacy Error Rate (WAER) following TER (Snover et al., 2006):

$$\texttt{WAER} = \frac{w_1\#\texttt{W} + w_2\#\texttt{WT} + w_3\#\texttt{M} + w_4\#\texttt{MT}}{\texttt{Len(S)}} \tag{1}$$

where #W, #WT, #M, #MT are the number of wrong words, wrong terminologies, missing words, missing terminologies respectively; w_1, w_2, w_3, w_4 are corresponding weight for each error class, indicating its importance. Len(S) is the source sentence length.

With only missing and wrong error class, WAER is not sufficient to serve as an independent eval-

	Base	+ pred.	+ gold.
BLEU	0.482	0.524	0.541
METEOR	0.638	0.673	0.709
ROUGE-L	0.582	0.645	0.699
BEER	0.582	0.640	0.681
CUNI-TreeAggreg	0.535	0.625	0.670
MEANT_2.0	0.639	0.641	0.680
MEANT_2.0-nosrl	0.630	0.633	0.672
UHH_TSKM	0.477	0.600	0.657
BLEU2VEC_sep	0.526	0.588	0.620
CHRF	0.591	0.641	0.685
CHRF++	0.593	0.644	0.690
NGRAM2VEC	0.520	0.576	0.607
Ave	0.573	0.622	0.662

Table 5: Segment-level Pearson correlation of various metrics and their enhancements with our $WAER$ measure through (Eq. 2). + pred./+gold means $WAER$ is calculated by our error detection model predicted results/gold standard as per se.

uation metric. As a common practice in BLEU and METEOR, we view WAER as a penalty to base metrics, which we refer as Score, and calculate the penalized score as follow:

$$\widetilde{\texttt{Score}} = \texttt{Score} * (1 - \texttt{WAER}) \qquad (2)$$

To determine w_1, w_2, w_3, w_4, we build a development set of 300 instances, sampled from our *TransErr*, with overall translation quality scored by multiple judges who are professional at both English and Chinese. To avoid bias, our error annotators of *TransErr* are excluded from this scoring. We adjust w_1, w_2, w_3, w_4 in this development set based on our gold annotation, such that the enhanced metrics achieves highest correlation with the average human score.

We examine three benchmark metrics (BLEU, METEOR, ROUGE) and the recent metrics submitted to WMT17: BEER (Stanojević and Sima'an, 2015), CUNI-TreeAggreg (Mareček et al., 2017), MEANT_2.0, MEANT_2.0-nosrl (Lo, 2017), UHH_TSKM (Duma and Menzel, 2017), NGRAM2VEC, BLEU2VEC_sep (Tättar and Fishel, 2017), CHRF (Popović, 2015), CHRF++ (Popović, 2017). We test their performance on WMT 17 DA dataset. The results are in Table 5. WAER is calculated by both system prediction results and the gold standard. We observe that, with our error prediction, WAER is able to enhance almost all metrics substantially (+0.049 on average). When calculating WAER using the gold standard, the performance is even higher (+0.089 on average). This demonstrates the efficacy of error detection for enhancing existing direct assessment metrics. Note that, we do not conclude that our WAER (Eq. 1) and the way of enhancing existing Score (Eq. 2) is optimal in terms of achieving better correlations with human judgment. We also want to find out which error class affects the overall translation quality more. Therefore, we study how the enhanced metrics correlate to human judgment when the weight of a single class of error changes. The trend is shown in Figure 4 where y-axis is the average correlation of all metrics enhanced by WAER on gold standard. X-axis is the weight of the observed error class. To eliminate the contribution made by other error classes, weights of other error classes are set to 0. From Figure 4, one can observe that wrong related error (W and WT) can help existing metrics achieve higher correlation to human judgment than missing-related error (M and MT). This suggests wrong translation is more severe than missing word. This provides valuable feedback on the advancement of evaluation metric towards human evaluation. Similarly, one can also observe that adequacy issue on terminologies is more severe than normal words and term-related errors need higher weights to achieve highest correlation. Accurate translation of terminology is an issue in neural machine translation if its occurrence is low in the training data. Accurate translation of terminology may be an area that warrens more work.

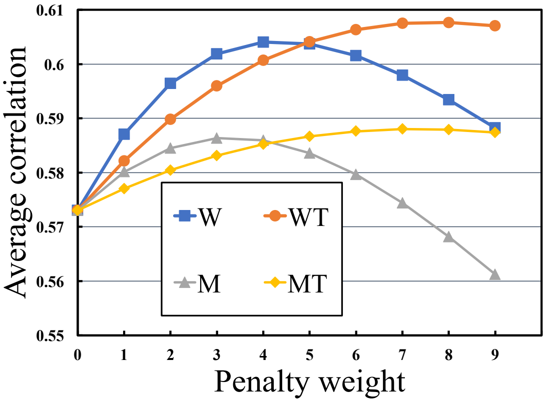

Figure 4: Correlation with respect to each error class.

7 Conclusion

In this paper, we revive the problem of error detection task for fine-grained machine translation evaluation purpose. We contribute a high-quality dataset *TransErr* with *Missing* and *Wrong* translation manually annotated by professional translators to enable the development of supervised methods. Based on *TransErr*, we benchmark this task by proposing a strong baseline, i.e., *AlignDet* model, based on our *Alignment Triangle* observation. We also conduct various discussions about this task such as the challenges of unsupervised methods for error annotation, the bottleneck of *AlignDet* model and the potential benefits of the error detection task for existing MT evaluation metrics task.

This work represents the preliminary work for more multi-faceted machine translation evaluation, focusing on multiple aspects instead of only a score or ranking, with the goal of push MT techniques to a higher standard. However, there are areas for further exploration. In the future, we will explore more advanced techniques such as neural networks for error detection. We will also investigate how to make better evaluation metrics with the help of error detection. We will also study how to improve machine translation model through error detection. For example, how to improve MT models according to different class of errors.

8 Acknowledgements

NExT++ research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative. The work is also supported by The Translational RD Grant (TRANS Grant) initiative managed by the Government Technology Agency of Singapore (GovTech) and the National Research Foundation (NRF), Singapore.

References

Ondřej Bojar, Yvette Graham, and Amir Kamran. 2017. Results of the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 489–513, Copenhagen, Denmark. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 10–51, Montréal, Canada. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Melania Duma and Wolfgang Menzel. 2017. Uhh submission to the wmt17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 582–588, Copenhagen, Denmark. Association for Computational Linguistics.

Kai Fan, Bo Li, Fengming Zhou, and Jiayi Wang. 2018. " bilingual expert" can find translation errors. *arXiv preprint arXiv:1807.09433*.

Mark Fishel, Ondrej Bojar, and Maja Popovic. 2012. Terra: a collection of translation error-annotated corpora. In *LREC*, pages 7–14.

Francisco Guzmán, Shafiq Joty, Lluís Màrquez, and Preslav Nakov. 2014. Using discourse structure improves machine translation evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 687–698.

Aaron Li-Feng Han, Yi Lu, Derek F. Wong, Lidia S. Chao, Liangye He, and Junwen Xing. 2013. Quality estimation for machine translation using the joint method of evaluation criteria and statistical modeling. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 365–372, Sofia, Bulgaria. Association for Computational Linguistics.

Junjie Hu, Wei-Cheng Chang, Yuexin Wu, and Graham Neubig. 2018. Contextual encoding for translation quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 788–793, Belgium, Brussels. Association for Computational Linguistics.

Julia Ive, Frédéric Blain, and Lucia Specia. 2018. Deepquest: a framework for neural-based quality estimation. *In the Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics, Sante Fe, New Mexico, USA.*

Hyun Kim, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator using multilevel task learning with stack propagation for neural quality estimation. pages 562–568.

Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics*, 35(4):505–512.

Ding Liu and Daniel Gildea. 2005. Syntactic features for evaluation of machine translation. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 25–32.

Yang Liu and Maosong Sun. 2015. Contrastive unsupervised word alignment with non-local features. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Chi-kiu Lo. 2017. Meant 2.0: Accurate semantic mt evaluation for any output language. In *Proceedings of the second conference on machine translation*, pages 589–597.

Ngoc Quang Luong, Benjamin Lecouteux, and Laurent Besacier. 2013. LIG system for WMT13 QE task: Investigating the usefulness of features in word confidence estimation for MT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 386–391, Sofia, Bulgaria. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

David Mareček, Ondřej Bojar, Ondřej Hübsch, Rudolf Rosa, and Dusan Varis. 2017. Cuni experiments for wmt17 metrics task. In *Proceedings of the Second Conference on Machine Translation*, pages 604–611.

F. J. Och and H. Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Evaluating machine translation with lfg dependencies. *Machine Translation*, 21(2):95–119.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 425–430.

Maja Popović. 2011b. Hjerson: An open source tool for automatic error classification of machine translation output. *The Prague Bulletin of Mathematical Linguistics*, 96:59–67.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović and Mihael Arcan. 2016. Pe2rr corpus: manual error annotation of automatically pre-annotated mt post-edits. In *Proceedings of the tenth international conference on Language Resources and Evaluation (LREC 2016)*, pages 27–32.

Maja Popović, Aljoscha Burchardt, et al. 2011a. From human to automatic error classification for machine translation output. In *15th International Conference of the European Association for Machine Translation (EAMT 11)*.

Maja Popović and Hermann Ney. 2011. Towards automatic error analysis of machine translation output. *Computational Linguistics*, 37(4):657–688.

Darcey Riley and Daniel Gildea. 2010. Improving the performance of giza++ using variational bayes. *The University of Rochester, Computer Science Department, Tech. Rep*, 963.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. Ruse: Regressor using sentence embeddings for automatic machine translation evaluation. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 764–771, Belgium, Brussels. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or hter?: exploring different human judgments with a tunable mt metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268. Association for Computational Linguistics.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón Astudillo, and André FT Martins. 2018. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 689–709.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1):39–50.

Miloš Stanojević and Khalil Sima'an. 2015. Beer 1.1: Illc uva submission to metrics and tuning task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 396–401.

Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2014. Back to basics for monolingual alignment: Exploiting word similarity and contextual evidence. *Transactions of the Association for Computational Linguistics*, 2:219–230.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2014. Recurrent neural networks for word alignment model. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1470–1480.

Andre Tättar and Mark Fishel. 2017. bleu2vec: the painfully familiar metric on continuous vector space steroids. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 619–622, Copenhagen, Denmark. Association for Computational Linguistics.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated dp based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

David Vilar, Jia Xu, D'Haro Luis Fernando, and Hermann Ney. 2006. Error analysis of statistical machine translation output. In *LREC*, pages 697–702.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the 16th conference on Computational linguistics-Volume 2*, pages 836–841. Association for Computational Linguistics.

Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for wmt18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 809–815.

Guillaume Wisniewski, Natalie Kübler, and François Yvon. 2014. A corpus of machine translation errors extracted from translation students exercises. In *LREC*, pages 3585–3588.

Daniel Zeman, Mark Fishel, Jan Berka, and Ondřej Bojar. 2011. Addicter: what is wrong with my translations? *The Prague Bulletin of Mathematical Linguistics*, 96:79–88.

Zaixiang Zheng, Shujian Huang, Zhaopeng Tu, Xin-Yu Dai, and Jiajun Chen. 2019. Dynamic past and future for neural machine translation.