Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan*, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

Microsoft AI & Research

Abstract

Machine translation has made rapid advances in recent years. Millions of people are using it today in online translation systems and mobile applications in order to communicate across language barriers. The question naturally arises whether such systems can approach or achieve parity with human translations. In this paper, we first address the problem of how to define and accurately measure human parity in translation. We then describe Microsoft's machine translation system and measure the quality of its translations on the widely used WMT 2017 news translation task from Chinese to English. We find that our latest neural machine translation system has reached a new state-of-the-art, and that the translation quality is at human parity when compared to professional human translations. We also find that it significantly exceeds the quality of crowd-sourced non-professional translations.

1 Introduction

Recent years have seen human performance levels reached or surpassed in tasks ranging from games such as Go [32] to classification of images in ImageNet [20] to conversational speech recognition on the Switchboard task [49].

In the area of machine translation, we have seen dramatic improvements in quality with the advent of attentional encoder-decoder neural networks [34, 3, 38]. However, translation quality continues to vary a great deal across language pairs, domains, and genres, more or less in direct relationship to the availability of training data. This paper summarizes how we achieved human parity in translating text in the news domain, from Chinese to English. While the techniques we used are not specific to the news domain or the Chinese-English language pair, we do not claim that this result necessarily generalizes to other language pairs and domains, especially where limited by the availability of data and resources.

Translation of news text has been an area of active interest in the Machine Translation community for over a decade, due to the practical and commercial importance of this domain, the availability of abundant parallel data on the web (at least in the most popular languages) and a long history of government-funded projects and evaluation campaigns, such as NIST-OpenMT[1] and GALE[2]. The annual evaluation campaign of the WMT (Conference on Machine Translation) [6], has also focused on news translation for more than a decade.

Defining and measuring human quality in translation is challenging for a number of reasons. Traditional metrics of translation quality, such as BLEU [28], TER [33] and Meteor [10] measure translation quality by comparison with one or more human reference translations. However, the same source sentence can be translated in sometimes substantially different but equally correct ways. This makes reference-based evaluation nearly useless in determining quality of human translations or near-human-quality machine translations.

*Corresponding author: hanyh@microsoft.com

[1] https://www.nist.gov/itl/iad/mig/open-machine-translation-evaluation

[2] https://www.nist.gov/itl/iad/mig/machine-translation-evaluation-gale

Further complicating matters, we find that the quality of reference translations, long assumed to be "gold" annotations by professional translators, are sometimes of remarkably poor quality. This is because references are often crowd-sourced (either directly, or indirectly through translation vendors). We have observed that crowd workers often use on-line MT with or without post-editing, rather than translating from scratch. Furthermore, many crowd workers appear to have only a rudimentary grasp of one of the languages, which often leads to unacceptable translation quality.

In Section 2, we describe how we address these challenges in defining and measuring human quality. In Section 3, we describe our system architecture. Section 4 describes our data and experiments. Sections 5 and 6 present our evaluation results and analysis.

2 Human Parity on Translation

Achieving human parity for machine translation is an important milestone of machine translation research. However, the idea of computers achieving human quality level is generally considered unattainable and triggers negative reactions from the research community and end users alike. This is understandable, as previous similar announcements have turned out to be overly optimistic.

Before any meaningful discussion of human parity can occur, we require a rigorous definition of the concept of human parity for translation. Based on this theoretical definition we can then investigate how close neural machine translation is to this goal.

2.1 Defining Human Parity

Intuitively, we can define human parity for translation as follows:

Definition 1. *If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved* HUMAN PARITY.

Assuming that it is possible for humans to measure translation quality by assigning scores to translations of individual sentences of a test set, and generalizing from a single sentence to a set of test sentences, this effectively yields the following statistical definition:

Definition 2. *If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved* HUMAN PARITY.

We choose definition 2 to address the question of human parity for machine translation in a fair and principled way. Given a reliable scoring metric to determine translation quality, based on direct human assessment, one can use a paired statistical significance test to decide whether a given machine translation system can be considered at parity with human translation quality for a test set and corresponding human references.

It is important to note that this definition of human parity does not imply that the machine translation system *outperforms* the human benchmark, but rather that its quality is statistically *indistinguishable*. It also does not imply that the translation is error-free. Machines, like humans, will continue to make mistakes.

Finally, achieving human parity on a given test set is measured with respect to a specific set of benchmark human translations and does not automatically generalize to other domains or language pairs.

2.2 Judging Human Parity

Our operational definition of human parity requires that human annotators be used to judge translation quality. While there exist various automated metrics to measure machine translation quality, these can only act as a (not necessarily correlated) proxy. Such metrics are typically reference-based and thus subject to *reference bias*. This can occur in the form of bad reference translations which result in bad segment scores. Also, due to the generative nature of translation, there often are multiple valid translations for a given input segment. Any translation which does not closely match the structure of the corresponding reference has a scoring disadvantage, even

perfect human translations. While these effects can be lessened using multiple references, the underlying problem remains unsolved[3].

Therefore, following the Conference on Machine Translation (WMT) we adopt *direct assessment* described in WMT17 [6] as our human evaluation method. To avoid reference bias—which can also happen for human evaluation[4]—we use the *source-based* evaluation methodology following IWSLT17 [7].

In source-based *direct assessment*, annotators are shown source text and a candidate translation and are asked the question *"How accurately does the above candidate text convey the semantics of the source text?"*, answering this using a slider ranging from 0 (*Not at all*) to 100 (*Perfectly*).[5] As a side effect, we have to employ bilingual annotators for our human evaluation campaigns.

The raw human scores are then standardized to a z-score, defined as the signed number of standard deviations an observation is above the mean, relative to a sample.

The z-scores are then averaged at the segment and system level. Results with statistically insignificant differences are grouped into clusters (according to Wilcoxon rank sum test [44] at p-level $p \leq 0.05$).[6]

To identify unreliable crowd workers, direct assessment includes artificially degraded translation output, so called "bad references". Any large scale crowd annotation task requires such integrated quality controls to guarantee high quality results. In our evaluation campaigns for Chinese into English, we observed relatively few attempts of gaming or spamming compared to other languages for which we run similar annotation tasks (we do not report on those in the context of this paper). In the remainder of this paper, direct assessment ranking clusters are computed in the same way as they had been generated for the WMT17 conference, with minor modifications[6].

3 System Description

3.1 Neural Machine Translation

Neural Machine Translation (NMT) [3] represents the state-of-the-art for translation quality. This has been demonstrated in various research evaluation campaigns (e.g. WMT [6]), and also for large scale production systems [45, 11]. NMT scales to train on parallel data on the order of tens of millions of sentences.

Currently, State-of-the-art NMT [3, 34] is generally based on a sequence-to-sequence encoder-decoder model with an attention mechanism [3]. Attentional sequence-to-sequence NMT models the conditional probability $p(\mathbf{y}|\mathbf{x})$ of the translated sequence $\mathbf{y}$ given an input sequence $\mathbf{x}$. In general, an attentional NMT system θ consists of two components: an encoder θ_e which transforms the input sequence into a sequence or set of continuous representations, and a decoder θ_d that dynamically reads out the encoder's output with an attention mechanism and predicts the conditional distribution of each target word. Generally, θ is trained to maximize the likelihood on a parallel training set consisting of N sentence pairs:

$$\begin{aligned}\mathcal{L}(\theta) &= \sum_{n=1}^{N} \log p\left(\mathbf{y}^{(n)}|\mathbf{x}^{(n)};\theta\right) \\ &= \sum_{n=1}^{N}\sum_{t=1}^{T} \log p\left(y_t^{(n)}|\mathbf{y}_{<t}^{(n)}, h_{t-1}^{(n)}, f^{\text{att}}\left(f^{\text{enc}}\left(\mathbf{x}^{(n)}\right), \mathbf{y}_{<t}^{(n)}, h_{t-1}^{(n)},\right);\theta\right) \end{aligned} \quad (1)$$

where $h_{t-1}^{(n)}$ denotes an internal decoder state, and $\mathbf{y}_{<t}$ the words preceding step t. At each step t, the attention mechanism f^{att} determines a context vector as a weighted sum over the outputs

[3] HyTER [12] attempted to solve this but did not achieve mainstream success.

[4] Results from both source-based and reference-based direct assessment collected for IWSLT17 [7] show that annotators assign higher scores in the source-based scenario and that they are more strict with their scoring in the reference-based scenario. This indicates that references do in fact influence human scoring behavior. Consequently, bad references will affect human evaluation in a reference-based direct assessment.

[5] Co-author Christian Federmann, in his role as co-organizer of the annual WMT evaluation campaign, was instrumental in developing the Appraise evaluation system used by WMT and also in this paper. He was not involved in developing the systems being evaluated here, nor were the human benchmark references available to the system developers. Hence, our evaluation was implemented in a double-blind manner.

[6] WMT17 implemented this using R's `wilcox.test()`. Our implementation differs from this as the clustering has been integrated into Appraise and uses the Mann-Whitney rank test [26] at the same p-level $p \leq 0.05$, based on Python's `scipy.mannwhitneyu()`. For the purpose of determining if the difference between scores for two candidate systems is statistically significant, both implementations are equivalent.

of the encoder $f^{\text{enc}}\left(\mathbf{x}^{(n)}\right)$, where the weights are determined essentially by comparing each of the encoder's outputs against the decoder's internal state and output up to time $t-1$. f^{enc} is a sentence-level feature extractor and can be implemented as multi-layer bidirectional RNNs [3, 45], a convolutional model (ConvS2S), [16] or a Transformer [38].

Like RNN sequence-to-sequence models, ConvS2S and Transformer utilize an encoder-decoder architecture. However, both models aim to eliminate the internal decoder state h_{t-1}. This side steps the recurrent nature of RNN, in which each sentence is encoded word by word, which limits the parallelizability of the computation and makes the encoded representation sensitive to the sequence length.

ConvS2S utilizes a stacked convolutional representation that models the dependencies between nearby words on lower layers, while longer-range dependencies are handled in the upper layers of the stack. The decoder applies attention on each layers. ConvS2S also utilizes position sensitive embeddings along with residual connections to accommodate positional variance.

The Transformer model replaces the convolutions with self-attention, which also eliminates the recurrent processing and positional dependency in the encoder. It also utilizes multi-head attention, which allows to attend to multiple source positions at once, in order to model different types of dependencies regardless of position. Similar to ConvS2S, the Transformer model utilizes positional embeddings to compensate for the ordering information, though it proposes a non-parametric representation. While these models eliminate recurrence in the encoder, all models discussed above decode auto-regressively, where each output word's distribution is conditioned on previously generated outputs. The Transformer model has shown [38] to yield significant improvement and therefore was choses as the base for our work in this paper.

3.2 Reaching Human Parity

Despite immense progress on NMT in the research community over the past years, human parity has remained out of reach. In this paper, we describe our efforts to achieve human parity on large-scale datasets for a Chinese-English news translation task. We address a number of limitations of the current NMT paradigm. Our contributions are:

- We utilize the duality of the translation problem to allow the model to learn from both source-to-target and target-to-source translations. Simultaneously this allows us to learn from both supervised and unsupervised source and target data. This will be described in Section 3.3. Specifically, we utilize a generic Dual Learning approach [19, 47, 46] (Section 3.3.1), and introduce a joint training algorithm to enhance the effect of monolingual source and target data by iteratively boosting the source-to-target and target-to-source translation models in a unified framework (Section 3.3.2).

- NMT systems decode auto-regressively from left-to-right, which means that during sequential generation of the output, previous errors will be amplified and may mislead subsequent generation. This is only partially remedied by beam search. We propose two approaches to alleviate this problem: Deliberation Networks [48] is a method to refine the translation based on two-pass decoding (Section 3.4.1); and a new training objective over two Kullback-Leibler (KL) divergence regularization terms encourages agreement between left-to-right and right-to-left decoding results (Section 3.4.2).

- Since NMT is very vulnerable to noisy training data, rare occurrences in the data, and the training data quality in general [4]. We discuss our approaches for data selection and filtering, including a cross-lingual sentence representation, in Section 3.5.

- Finally, we find that our systems are quite complementary, and can therefore benefit greatly from system combination, ultimately attaining human parity. See section 3.6.

In this work, we interchangeably use source-to-target and (Zh→En) to denote Chinese-to-English; target-to-source and (En→Zh) to denote English-to-Chinese.

3.3 Exploiting the Dual Nature of Translation

We leverage the duality of the translation problem to allow the model to learn from both source-to-target and target-to-source translations. We explore the translation duality using two approaches: Dual Learning 3.3.1 and Joint Training 3.3.2

3.3.1 Dual Learning for NMT

Dual learning [19, 47, 46], a recently proposed learning paradigm, tries to achieve the co-growth of machine learning models in two dual tasks, such as image classification vs. image generation, speech recognition vs. text-to-speech, and Chinese to English vs. English to Chinese translation. In dual learning, the two parallel models (referred to as the *primal model* and the *dual model*) enhance each other by leveraging primal-dual structure in order to learn from unlabeled data or regularize the learning from labeled data. Ever since dual learning was proposed, it has been successfully applied to various real-world problems such as question answering [35], image classification [47], image segmentation [25], image to image translation [50, 52, 24], face attribute manipulation [31], and machine translation [19, 43, 23, 1].

In this work, to achieve strong machine translation performance, we combine two different dual learning methods that respectively enhance the usage of monolingual and bilingual training data. We set the Chinese to English (Zh→En) translation model as the primal model and the English to Chinese (En→Zh) model as the dual model, respectively denoted as $p(\mathbf{y}|\mathbf{x};\theta_{x\to y})$ and $p(\mathbf{x}|\mathbf{y};\theta_{y\to x})$.

- *Dual unsupervised learning* (DUL) [19]. To enhance the Zh→En translation quality, DUL efficiently leverages a monolingual Chinese corpus based on additional supervision signals from the dual En→Zh model. Concretely speaking, for a monolingual Chinese sentence $\mathbf{x}$, an English translation $\mathbf{y}$ is sampled using the primal model $p(\cdot|\mathbf{x};\theta_{x\to y})$; starting from $\mathbf{y}$, we use the dual model $p(\cdot|\mathbf{y};\theta_{y\to x})$ to compute the log-likelihood $\log p(\mathbf{x}|\mathbf{y};\theta_{y\to x})$ of reconstructing $\mathbf{x}$ from $\mathbf{y}$ and treat it as the reward of taking action $\mathbf{y}$ at state $\mathbf{x}$. We would like to maximize the expected reconstruction log-likelihood when iterating over all possible translation $\mathbf{y}$ for $\mathbf{x}$, shown as:

$$\mathcal{L}(\mathbf{x};\theta_{x\to y}) = E_{\mathbf{y}\sim p(\cdot|\mathbf{x};\theta_{x\to y})}\big\{\log p(\mathbf{x}|\mathbf{y};\theta_{y\to x})\big\} = \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x};\theta_{x\to y}) \log p(\mathbf{x}|\mathbf{y};\theta_{y\to x}) \tag{2}$$

  Taking the gradient of $\mathcal{L}(\mathbf{x};\theta_{x\to y})$ with respect to $\theta_{x\to y}$, we obtain:

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathbf{x};\theta_{x\to y})}{\partial \theta_{x\to y}} &= \sum_{\mathbf{y}} \frac{\partial p(\mathbf{y}|\mathbf{x};\theta_{x\to y})}{\partial \theta_{x\to y}} \log p(\mathbf{x}|\mathbf{y};\theta_{y\to x}) \\ &= \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x};\theta_{x\to y}) \frac{\partial \log p(\mathbf{y}|\mathbf{x};\theta_{x\to y})}{\partial \theta_{x\to y}} \log p(\mathbf{x}|\mathbf{y};\theta_{y\to x})\end{aligned} \tag{3}$$

  Since summing over all possible $\mathbf{y}$ in the above equation is computationally intractable, we use Monte Carlo sampling to approximate the above expectation:

$$\frac{\partial \mathcal{L}(\mathbf{x};\theta_{x\to y})}{\partial \theta_{x\to y}} \approx \frac{\partial \log p(\mathbf{y}'|\mathbf{x};\theta_{x\to y})}{\partial \theta_{x\to y}} \log p(\mathbf{x}|\mathbf{y}';\theta_{y\to x}), \tag{4}$$

  where $\mathbf{y}'$ is a sampled translation from the primal model $p(\cdot|\mathbf{x};\theta_{x\to y})$.

  The approximated gradient is used to update the primal model parameters $\theta_{x\to y}$. Note that the parameters of the dual model $\theta_{y\to x}$ can be updated using a monolingual English corpus in a similar way by maximizing the reconstruction likelihood from possible Chinese translations.

- *Dual supervised learning* (DSL) [47]. Unlike DUL, which aims to effectively leverage monolingual data, DSL is an approach to better utilize bilingual training data by enhancing probabilistic correlations within the two models. The idea of DSL is to force the joint probability consistency within primal model and dual model. Specifically, for a bilingual sentence pair $(\mathbf{x},\mathbf{y})$, ideally we have $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$. However, if the two models are trained separately, it is hard for them to satisfy $p(\mathbf{x})p(\mathbf{y}|\mathbf{x}) = p(\mathbf{y})p(\mathbf{x}|\mathbf{y})$. Therefore, when applied in neural machine translation, DSL conducts joint training of the two models and introduces an additional loss term on the parallel data $(\mathbf{x},\mathbf{y})$ for regularization:

$$\mathcal{L}_{DSL} = (\log \hat{p}(\mathbf{x}) + \log p(\mathbf{y}|\mathbf{x};\theta_{x\to y}) - \log \hat{p}(\mathbf{y}) - \log p(\mathbf{x}|\mathbf{y};\theta_{y\to x}))^2, \tag{5}$$

  where $\hat{p}(\mathbf{x})$ and $\hat{p}(\mathbf{y})$ are empirical marginal distributions induced by the training data. In our experiments, they are the output scores of two language models respectively trained on Chinese and English corpus containing both bilingual and monolingual data.

In our architecture, both DUL and DSL are used in model training, both of which are applied to the monolingual and bilingual training corpora.

3.3.2 Joint Training of Source-to-Target and Target-to-Source Models

Back translation [29] augments relatively scarce parallel data with plentiful monolingual data, allowing us to train source-to-target (S2T) models with the help of target-to-source (T2S) models. Specifically, given a set of sentences $\{\mathbf{y}^{(t)}\}$ in the target language, a pre-constructed T2S translation system is used to automatically generate translations $\{\mathbf{x}^{(t)}\}$ in the source language. These synthetic sentence pairs $\{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}$ are combined with the original bilingual data when training the S2T NMT model. In order to leverage both source and target language monolingual data, and also let S2T and T2S models help each other, we leverage the joint training method described in [51] to optimize them by extending the back-translation method. The joint training method uses the monolingual data and updates NMT models through several iterations.

Given parallel corpus $D = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ and target monolingual corpus $Y = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$, a semi-supervised training objective is used to jointly maximize the likelihood of both bilingual data and monolingual data:

$$\mathcal{L}^*(\theta_{x\to y}) = \sum_{n=1}^{N} \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}) + \sum_{t=1}^{T} \log p(\mathbf{y}^{(t)}) \tag{6}$$

By introducing $\mathbf{x}$ as the latent variable representing the source translation of target sentence $\mathbf{y}^{(t)}$, Equation 6 can be optimized in an EM framework, with the help of a T2S translation model:

$$\mathcal{L}(\theta_{x\to y}) = \sum_{n=1}^{N} \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}) + \sum_{t=1}^{T} \sum_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}^{(t)}) \log p(\mathbf{y}^{(t)}|\mathbf{x}) \tag{7}$$

Similarly, we can optimize the T2S translation model with the help of S2T translation model as follows:

$$\mathcal{L}(\theta_{y\to x}) = \sum_{n=1}^{N} \log p(\mathbf{x}^{(n)}|\mathbf{y}^{(n)}) + \sum_{s=1}^{S} \sum_{\mathbf{y}} p(\mathbf{y}|\mathbf{x}^{(s)}) \log p(\mathbf{x}^{(s)}|\mathbf{y}) \tag{8}$$

As we can find from Equation 7 and 8, model $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{y})$ serve as each other's pseudo-training data generator: $p(\mathbf{x}|\mathbf{y})$ is used to translate Y into X for $p(\mathbf{y}|\mathbf{x})$, while $p(\mathbf{y}|\mathbf{x})$ is used to translate X to Y for $p(\mathbf{x}|\mathbf{y})$. The joint training process is illustrated in Figure 1. Before the first iteration starts, two initial translation models $p_0(\mathbf{y}|\mathbf{x})$ and $p_0(\mathbf{x}|\mathbf{y})$ are pre-trained with parallel data $D = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}$. This step is denoted as iteration 0 for sake of consistency. In iteration 1, two NMT systems $p_0(\mathbf{y}|\mathbf{x})$ and $p_0(\mathbf{x}|\mathbf{y})$ are used to translate monolingual data $X = \{\mathbf{x}^{(s)}\}$ and $Y = \{\mathbf{y}^{(t)}\}$, which creates two synthetic training data sets $X' = \{\mathbf{x}^{(s)}, \mathbf{y}_0^{(s)}\}$ and $Y' = \{\mathbf{y}^{(t)}, \mathbf{x}_0^{(t)}\}$. Models $p_1(\mathbf{y}|\mathbf{x})$ and $p_1(\mathbf{x}|\mathbf{y})$ are then trained on this augmented training data by combining Y' and X' with parallel data D. It is worth noting that n-best translations are used, and the selected translations are weighted with the translation probabilities given by the NMT model, so that the negative impact of noisy translations can be minimized. In iteration 2, the above process is repeated, and the synthetic training data are re-generated with the updated NMT models $p_1(\mathbf{y}|\mathbf{x})$ and $p_1(\mathbf{x}|\mathbf{y})$, which are presumably more accurate. The learned NMT models $p_2(\mathbf{y}|\mathbf{x})$ and $p_2(\mathbf{x}|\mathbf{y})$ are also expected to improve with better pseudo-training data. The training process continues until the performance on a development data set is no longer improved.

3.4 Beyond the Left-to-Right Bias

Current NMT systems suffer from the *exposure bias* problem [5]. Exposure bias refers to the problem that during sequential generation of output, previous errors will be amplified and mislead subsequent generation. We address this limitation in two ways: a two-pass decoding (Deliberation Networks) 3.4.1 and Agreement Regularization 3.4.2.

3.4.1 Deliberation Networks

Classical neural machine translation models generate a translation word by word from left to right, all in one pass. This is very different from human behavior such as, for instance, while writing

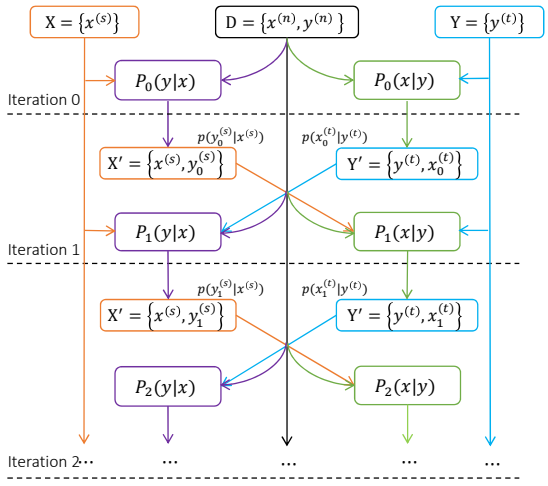

Figure 1: Illustration of joint training: S2T $p(\mathbf{y}|\mathbf{x})$ and T2S $p(\mathbf{x}|\mathbf{y})$

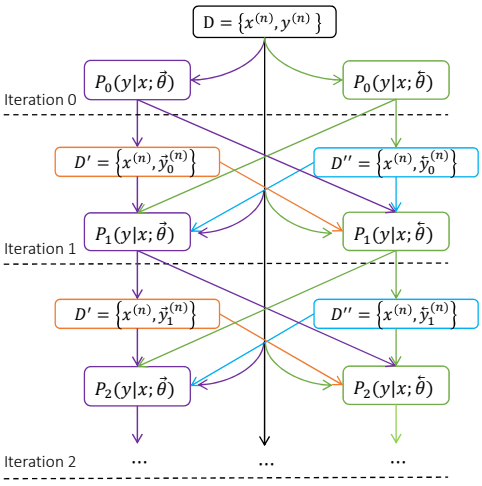

Figure 2: Illustration of agreement regularization: L2R $p(\mathbf{y}|\mathbf{x};\overrightarrow{\theta})$ and R2L $p(\mathbf{y}|\mathbf{x};\overleftarrow{\theta})$

articles or papers. When writing papers, usually we create a first draft, then we revisit the draft in its full context, further polishing each word (or phrase/sentence/paragraph) based on both its left-side context and right-side context. In contrast, in neural machine translation, decoding in only one pass makes the output of the t-th word y_t dependent on the source-side sentence $\mathbf{x}$ and its left context only (i.e., already generated tokens $\{y_1, \cdots, y_{t-1}\}$), without any opportunity to look into the future. Inspired by the human writing process, Deliberation Networks [48] try to overcome this drawback by decoding using a two-pass process with two decoders as illustrated in Fig. 3. The first-pass decoder outputs an initial translation as a draft. The second-pass decoder polishes this draft into a final translation. The draft translation output from the first pass decoder contains global information that enlarges the receptive field of decoding each token y_t in the second-pass decoding process, and thus breaks the limitation of only looking to the left-hand side.

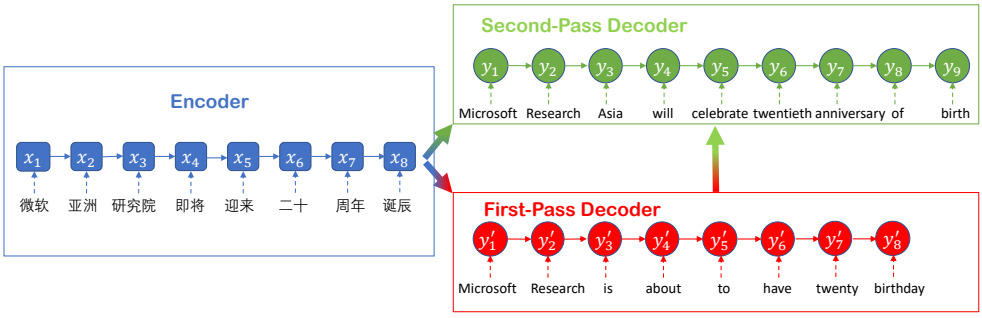

Figure 3: An example showing the decoding process of deliberation network.

The detailed model architecture, with a deliberation network built on top of Transformer, is shown in Fig. 4. As in standard Transformer, both the encoder $\mathcal{E}$ and the first-pass decoder $\mathcal{D}_1$ contain several stacked layers connected via a self attention mechanism. Specifically, the encoder assigns to each of the T_s source words a representation based on its original embedding and contextual information gathered from other positions. We denote this sequence of top-layer state vectors $h_{1:T_s}$ as $\mathcal{H}$. The encoder $\mathcal{E}$ reads the source sentence x and outputs a sequence of hidden states $\mathcal{H} = h_{1:T_s}$ via self attention. The first-pass decoder $\mathcal{D}_1$ takes $\mathcal{H}$ as inputs, conducts the first round decoding and obtains the first-pass translation sentence $\hat{\mathbf{y}}$ as well as the hidden states before softmax denoted as $\hat{\mathcal{S}}$. The second-pass decoder $\mathcal{D}_2$ also contains several stacked layers, but is significantly different from $\mathcal{D}_1$ in that $\mathcal{D}_2$ takes the hidden states output by both $\mathcal{E}$ and $\mathcal{D}_1$ as inputs. Specifically, denoting the output of the ith layer in $\mathcal{D}_2$ as s^i, we have $s^i = A_e(\mathcal{H}, s^{i-1}) + A_c(\hat{\mathcal{S}}, s^{i-1}) + A_s(s^{i-1})$, where A_e and A_c are the multi-head attention mechanism [38]

connecting $\mathcal{D}_2$ respectively with $\mathcal{E}$ and $\mathcal{D}_1$, and A_s is the self attention mechanism within $\mathcal{D}_2$ operating on s^{i-1}. It is easily observed that the last translation result y is dependent on the first translation sentence $\hat{\mathbf{y}}$, since we feed the outputs of the first-pass decoder $\mathcal{D}_1$ into the second-pass decoder $\mathcal{D}_2$. In this way we obtain global information on the target side, thereby allowing us to look at right context in sentence generation. Policy gradient algorithms are used to jointly optimize the parameters of the three parts.

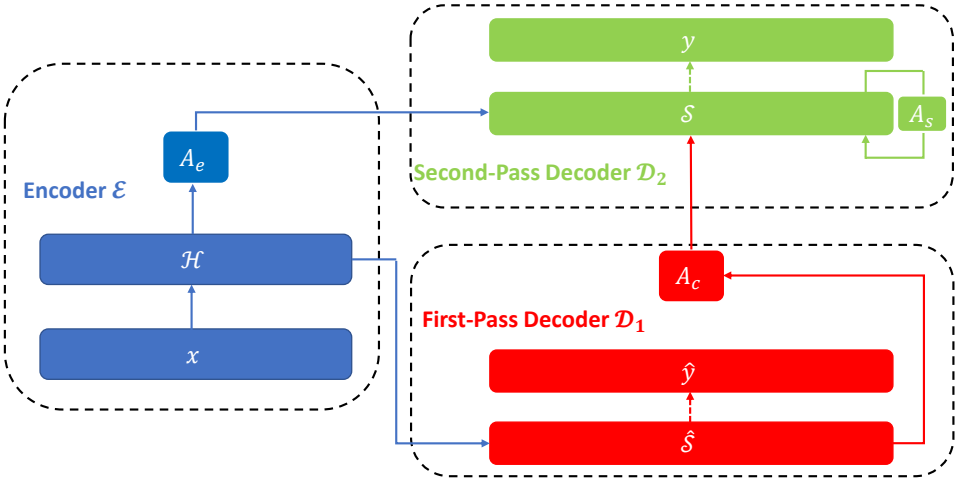

Figure 4: Deliberation network: Blue, red and green parts indicate encoder $\mathcal{E}$, first-pass decoder $\mathcal{D}_1$ and second-pass decoder $\mathcal{D}_2$ respectively. Solid lines represent the information flow via attention model. The self attention model within $\mathcal{E}$ and the $\mathcal{E}$-to-$\mathcal{D}_1$ attention model are omitted for readability.

The combination of dual learning and deliberation networks takes place as follows: First, we train the Zh→En and En→Zh Transformer models using both DUL and DSL. Then, for a target side monolingual sentence $\mathbf{y}$, the existing En→Zh model is used to translate it into Chinese sentence $\mathbf{x}'$. Afterwards, we treat (X', Y) as pseudo bilingual data and add it into the bilingual data corpus. The enlarged bilingual corpus is then used to train the deliberation network as described above. In deliberation network training, we use the Zh→En model obtained in the first step to initialize the encoder and first-pass decoder.

3.4.2 Agreement Regularization of Left-to-Right and Right-to-Left Models

An alternative way of addressing exposure bias is to leverage the fact that unsatisfactory translations with bad suffixes generated by a left-to-right (L2R) model usually have low prediction scores under a right-to-left (R2L) model. In the R2L model, if bad suffixes are fed as inputs to the decoder first, this will lead to corrupted hidden states, therefore good prefixes reached later will be given considerably lower prediction probabilities. This signal given by the R2L model can be leveraged to alleviate the exposure bias problem of the L2R model and vice versa.

To train the L2R model, two Kullback-Leibler (KL) divergence regularization terms are introduced into the maximum-likelihood training objective, as shown in

$$\begin{aligned}\mathcal{L}(\overrightarrow{\theta}) = &\sum_{n=1}^{N} \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \overrightarrow{\theta}) - \lambda \sum_{n=1}^{N} \mathrm{KL}(p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta})||p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})) \\ &- \lambda \sum_{n=1}^{N} \mathrm{KL}(p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})||p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta}))\end{aligned} \tag{9}$$

With a simple mathematic calculation and proper approximation, we can get the parameter gra-

Algorithm 1 Unified Joint Training Algorithm

Input: Bilingual Data $D = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$, Source and Target Monolingual Corpora $X = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$ and $Y = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$;

Output: S2T-L2R Model $p(\overrightarrow{\mathbf{y}}|\mathbf{x})$, S2T-R2L Model $p(\overleftarrow{\mathbf{y}}|\mathbf{x})$, T2S-L2R Model $p(\overrightarrow{\mathbf{x}}|\mathbf{y})$ and T2S-R2L Model $p(\overleftarrow{\mathbf{x}}|\mathbf{y})$;

procedure TRAINING PROCESS
  Pre-train four models with maximum likelihood on parallel corpora $D = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$;
  while Not Converged **do**
    Build weighted pseudo-parallel corpora $Y' = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^{T}$ with $p(\overrightarrow{\mathbf{x}}|\mathbf{y})$ using monolingual data $Y = \{\mathbf{y}^{(t)}\}_{t=1}^{T}$ as shown in Figure 1.
    Update $P(\overleftarrow{\mathbf{y}}|\mathbf{x})$ and $p(\overrightarrow{\mathbf{y}}|\mathbf{x})$ as shown in Figure 2, with original data $D = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ and synthetic data $Y' = \{(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})\}_{t=1}^{T}$.
    Build weighted pseudo-parallel corpora $\mathbf{x}' = \{(\mathbf{x}^{(s)}, \mathbf{y}^{(s)})\}_{s=1}^{S}$ with $p(\overrightarrow{\mathbf{y}}|\mathbf{x})$ using monolingual data $X = \{\mathbf{x}^{(s)}\}_{s=1}^{S}$ as introduced in Figure 1.
    Update $p(\overleftarrow{\mathbf{x}}|\mathbf{y})$ and $P(\overrightarrow{\mathbf{x}}|\mathbf{y})$ as shown in Figure 2, with original data $D = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ and synthetic data $X' = \{(\mathbf{x}^{(s)}, \mathbf{y}^{(s)})\}_{s=1}^{S}$.
  end while
end procedure

dients for L2R model as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}(\overrightarrow{\theta})}{\partial \overrightarrow{\theta}} = &\sum_{n=1}^{N} \frac{\partial \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} + \lambda \sum_{n=1}^{N} \sum_{\mathbf{y} \sim p(\cdot|\mathbf{x}^{(n)}; \overleftarrow{\theta})} \frac{\partial \log p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} \\ &+ \lambda \sum_{n=1}^{N} \sum_{\mathbf{y} \sim p(\cdot|\mathbf{x}^{(n)}; \overrightarrow{\theta})} \left(\log \frac{p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta})}{p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})} \frac{\partial \log p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})}{\partial \overrightarrow{\theta}} \right)\end{aligned} \tag{10}$$

The first part tries to maximize the log likelihood of the bilingual training corpus. The second part maximizes the log likelihood of the "pseudo corpus" constructed by the R2L model. The third part maximizes a weighted log likelihood of another pseudo corpus generated by the L2R model itself with a weight of $(\log(p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta})/p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})))$ which penalizes the samples where the L2R and R2L models do not agree. We find that the R2L model plays the role of an auxiliary system which provides a pseudo corpus in the second part and calculates the weight in the third part.

Similarly, we can get corresponding parameter gradients for the R2L model by introducing two KL divergence regularization terms, as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}(\overleftarrow{\theta})}{\partial \overleftarrow{\theta}} = &\sum_{n=1}^{N} \frac{\partial \log p(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}; \overleftarrow{\theta})}{\partial \overleftarrow{\theta}} + \lambda \sum_{n=1}^{N} \sum_{\mathbf{y} \sim p(\cdot|\mathbf{x}^{(n)}; \overrightarrow{\theta})} \frac{\partial \log p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta})}{\partial \overleftarrow{\theta}} \\ &+ \lambda \sum_{n=1}^{N} \sum_{\mathbf{y} \sim p(\cdot|\mathbf{x}^{(n)}; \overleftarrow{\theta})} \left(\log \frac{p(\mathbf{y}|\mathbf{x}^{(n)}; \overrightarrow{\theta})}{p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta})} \frac{\partial \log p(\mathbf{y}|\mathbf{x}^{(n)}; \overleftarrow{\theta})}{\partial \overleftarrow{\theta}} \right)\end{aligned} \tag{11}$$

With the help of the R2L model, the L2R model can be enhanced using Equation 10. With the enhanced L2R model, a better pseudo corpus and more accurate weights can be leveraged to improve the performance of the R2L model with Equation 11, while simultaneously this better R2L model can be reused to improve the L2R model. In such a way, L2R and R2L models can mutually boost each other as illustrated in Figure 2. The training process continues until the performance on a development data set is no further improving.

Since both the source and target sentences can be generated from left to right and from right to left, we can have a total of four systems, two source to target models: S2T-L2R (target sentence is generated from left to right), S2T-R2L (target sentence is generated from right to left), and two target to source models: T2S-L2R (source sentence is generated from left to right), T2S-R2L (source sentence is generated from right to left). Using the agreement regularization method described above, these four models can be optimized in a unified joint training framework, as shown in Algorithm 1. With the joint training method, a weighted pseudo corpus is generated by T2S-L2R model and used to train two S2T models (S2T-L2R and S2T-R2L) with the help of

agreement regularization. The enhanced S2T-L2R model is then used to build another weighted pseudo corpus to train two T2S models. These four systems boost each other until convergence is reached.

3.5 Data Selection and Filtering

Though NMT systems require huge amounts of training data, not all data are equally useful for training the systems. NMT systems are more vulnerable to noisy training data, rare occurrences in the data, and the training data quality in general. We are trying to tackle two different problems: selecting data relevant to the task and removing noisy data. Out-of-domain and noisy data are distinct problems and may harm the system in different ways. Many studies have highlighted the bad impact of noisy data on MT, such as [4]. Even small amounts of noisy data can have very bad effects since NMT models tend to assign high probabilities to rare events. Noise in data can take several forms, including totally incorrect translations, partial translations, inaccurate or machine translated data, wrong source or target language, or source copied to the target. We use features from word alignment to filter out the very noisy data, similar to the approach in [18]. However, data that is less egregiously noisy represents a bigger problem since it is harder to recognize.

The de-facto standard method for data selection for SMT is [27] and [2]. Unfortunately it has not proved as useful for NMT; while it reduces the training data it does not lead to improvements in system quality [37]. We propose a new approach that tackles both problems at once: filtering noisy data and selecting relevant data. Our approach centers on first learning a bilingual sentence vector representation where sentences in both languages are mapped into the same space. After learning this representation, we use it for both filtering noisy data and selecting relevant data.

To learn our sentence representation we train a unified bilingual NMT system similar to [53] that can translate between Chinese and English in both directions. We train this on a selected subset of the data that is known to be of good quality and in the relevant domain. Building the model with such relevant data has two advantages. First: it helps the representation to be similar to the cleaner data; second: relevant sentences would have better representation than irrelevant ones. Therefore we would achieve both data cleaning and relevant data selection objectives.

Recent progress in multi-lingual NMT i.e. [21] and [17] shows that these models are able to represent multiple languages in the same space. However, we don't use language markers because we want to force the model to learn similar representations for both Chinese and English. Given this bilingual system, for any sentence in Chinese or English we can run the encoder part of the system to get a contextual vector representation for each word of a sentence. This is the vector from the last encoder layer, normally used as input to the attention model. We represent each sentence vector as the mean of the word-level contextual vectors.

Specifically, the encoder assigns to each of the T_s source words a representation based on its original embedding and contextual information gathered from other positions. We denote this set of top-layer state vectors as $h_{1:T_s}$:

$$h_{1:T_s} = f^{\text{enc}}\big(E(x_1), ..., E(x_{T_s})\big) \tag{12}$$

where $E^I \in \mathbb{R}^{V \times d}$ is a look-up table of joint source and target embeddings, assigning each individual word a unique embedding vector.

If $h^{\text{enc}}_{1:T_s}$ denotes the encoder's top layer's output sequence, the sentence-vector representation S_{sv} of a given sentence S of length T_s is:

$$S_{sv} = \sum_{\ell=1}^{T_s} h^{\text{enc}}_{\ell} \tag{13}$$

A similarity measure SIM_{ST} between any two given sentences S and T, regardless of their languages, can be represented as the cosine similarity between their corresponding sentences vectors:

$$\text{SIM}_{ST} = \frac{S_{sv} \cdot T_{sv}}{|S_{sv}||T_{sv}|} \tag{14}$$

We train an RNN encoder-decoder system similar to [45] with 4 encoder layers with the first layer being bidirectional and 4 decoder layers and an attention model. After training the model, we run the encoder part only. Each resulting word context vector is composed of an 1024 dimension vector; therefore the sentence vector (S_{sv}) representation is of the same size.

For each sentence in the parallel training corpus, we measure the cross-lingual similarity between source and target sentences as in Equation 14. We reject sentences with similarity below a specified threshold. This approach enables us to drastically reduce the training data while significantly improving the accuracy. Since we use a model trained on relevant data, this data selection technique can serve a dual purpose by filtering noisy data as well as selecting relevant data.

3.6 System Combination and Re-ranking

In order to combine the systems described above, we combine n-best hypotheses from all systems and then train a re-ranker using k-best MIRA on the validation set. K-best MIRA [8] is a version of MIRA (a margin-based classification algorithm) that works with a batch tuning to learn a re-ranker for the k-best hypothesis.

The features we use for re-ranking are:

- SYS_{Score}: Original System Score and identifier.
- LM_{Score}: 5-gram language model trained on English news crawled data of 2015 and 2016.
- $R2L_{score}$: R2L system re-scoring. A system trained on Chinese source and reversed English target; the system is used to score each hypothesis.
- $E2Z_{score}$: English-to-Chinese system re-scoring. A system trained on English to Chinese is used to score each hypothesis. .
- ST_{SV} : Cross-lingual sentence similarity between source and the hypothesis as described in Section 3.5.
- $R2L_{SV}$: R2L sentence vector similarity: the best hypothesis from the R2L system is compared to each n-best hypothesis and used to generate a sentence similarity score based on sentence vector as above.
- $E2Z_{SV}$: Back Composition sentence vector similarity. A round trip translation is done for each n-best hypothesis to translate it back to Chinese. Then we use sentence vector similarity to measure the similarity between the original source and the recomposed source.

4 Experiments

In this section, we first introduce the data and experimental setup used in our experiments, and then evaluate each of the systems introduced in Section 3, both independently and after system combination and re-ranking.

4.1 Data and Experimental Setup

We use all of the available parallel data for the WMT17 Chinese-English translation task. This consists of about 332K sentence pairs from the News Commentary corpus, 15.8M sentence pairs from the UN Parallel Corpus, and 9M sentence pairs from the CWMT Corpus. We further filter the bilingual corpus according to the following criteria:

- Both the source and target sentences should contain at least 3 words and at most 70 words.
- Pairs where (source length $<$ 1.3$*$target length or target length $<$ 1.3$*$source length) are removed.
- Sentences with illegal characters (such as URLs, characters of other languages) are removed.
- Chinese sentences without any Chinese characters are removed.
- Duplicated sentence pairs are removed.

After filtration, we are left with 18M bilingual sentence pairs. We use the Chinese and English language models trained on the 18M sentences of bilingual data to filter the monolingual sentences from "News Crawl: articles from 2016" and "Common Crawl" provided by WMT17 using CED [27]. After filtering, we retain about 7M English and Chinese monolingual sentences. The monolingual data will be deployed in both dual learning and back-translation setups through the experiments.

Newsdev2017 is used as the development set and Newstest2017 as the test set. All the data (parallel and monolingual) have been tokenized and segmented into subword symbols using byte-pair encoding (BPE) [30]. The Chinese data has been tokenized using the Jieba tokenizer[7]. English sentences are tokenized using the scripts provided in Moses. We learn a BPE model with 32K merge operations, in which 44K and 33K sub-word tokens are adopted as source and target vocabularies separately.

4.2 Experimental Results

The Transformer model [39] is adopted as our baseline. Unless otherwise mentioned, all translation experiments use the following hyper-parameter settings based on Tensor2Tensor Transformer-big settings v1.3.0[8]. This corresponds to a 6-layer transformer with a model size of 1024, a feed forward network size (d_{ff}) of 4096, and 16 heads. All models are trained on 8 Tesla M40 GPUs for a total of 200K steps using the Adam [22] algorithm. The initial learning rate is set to 0.3 and decayed according to the "noam" schedule as described in [39].During training, the batch size is set to 5120 words per batch and checkpoints are created every 60 minutes. All results are reported on averaged parameters of the last 20 checkpoints. At test time, we use a beam of 8 and a length penalty of 1.0. All reported scores are computed using sacreBLEU v1.2.3,[9] which calculates tokenization-independent BLEU [28].[10]

The first section of Table 1 shows the results for the baselines. First we compare with the Sogou system [42], which was the best result reported at WMT 2017 evaluation campaign. Though Sogou is an ensemble of many systems, we reference it here for comparison. The rest of the systems reported in the table are single systems. Our baseline system, labeled **Base**, is trained on 18M sentences. **BT** is adding the back-translated data to the baseline.

SystemID	**Settings**	**BLEU**
Sogou	WMT 2017 best result [42]	26.40
Base	Transformer Baseline	24.2
BT	+Back Translation	25.57
DL	BT + Dual Learning	26.51
DLDN	BT + Dual Learning + Deliberation Nets	27.40
DLDN2	DLDN without first decoder reranking	27.20
DLDN3	BT+ Dual Learning + R2L sampling	26.88
DLDN4	BT+ Dual Learning + Bi-NMT	27.16
AR	BT + Agreement Regularization	26.91
ARJT	BT + Agreement Regularization + Joint Training	27.38
ARJT2	ARJT + dropout=0.1	27.19
ARJT3	ARJT + dropout=0.05	27.07
ARJT4	ARJT + dropout=0.01	26.98

Table 1: Automatic (BLEU) evaluation results on the WMT 2017 Chinese-English test set

Experimental Results of Dual Learning and Deliberation Networks

Our Dual Learning system consists of a Zh→En model and an En→Zh model, each adopting the same model configuration as the baseline (Base). For the deliberation network, the encoder and the first-pass decoder are initialized from the Zh→En model in the Dual Learning system,

[7] https://github.com/fxsjy/jieba
[8] https://github.com/tensorflow/tensor2tensor/blob/master/tensor2tensor/models/transformer.py
[9] https://github.com/awslabs/sockeye/tree/master/contrib/sacrebleu
[10] sacreBLEU signature: `BLEU+case.mixed+lang.zh-en+numrefs.1+smooth.exp_+test.wmt17/improved+tok.13a+version.1.2.3`

and the second pass decoder share the same model structures with the first-pass decoder. The evaluation results of the Dual Learning and Deliberation Network systems on WMT 2017 Chinese-English test set are listed in the second section of Table 1. Dual Learning makes more efficient use of the monolingual sentences and exploits the duality between Zh→En and En→Zh translation directions. Based on system **BT**, the Dual Learning system **DL** achieves 26.51 BLEU, a 0.94 point improvement over the **BT** system, and outperforms the best ensemble result of 26.40 in the WMT 2017 Chinese-English challenge . The Deliberation Network is further applied to the Dual Learning system, which is denoted as **DLDN**. The Deliberation Network aims to improve sentence generation quality by incorporating the global information provided by a first pass decoder. The **DLDN** system further achieves a BLEU score of 27.40, a 0.89 BLEU score improvement over the already strong **DL** system.

We also explore some variants of our **DL** and **DLDN** systems, denoted as **DLDN2/3/4** in the second section of Table 1. In **DLDN**, we use both the first and second pass decoders to rerank the generated sentence and choose the top-1 result. In system **DLDN2**, we then remove this reranking to see how the performance changes, yielding a 27.20 BLEU score, a 0.2 point drop. In system **DLDN3**, we replace the Deliberation Network with R2L sampling. R2L sampling is a data augmentation technique where we first train a Zh→En model that generates sentences in a right-to-left(R2L) manner by reversing the target sentence in the training data, and use the R2L model to sample English sentences given monolingual Chinese sentences. We can see that adding R2L sampling to Dual Learning indeed brings BLEU score improvements, but performs worse than the Deliberation Network. In system **DLDN4**, we further add Bi-NMT, which bidirectionally generates candidate sentences in a single model, on the **DL** system and achieve 27.16 BLEU score.

Experimental Results of Agreement Regularization and Joint Training

Data enhancement has been shown to improve NMT performance. We proposed the agreement regularization approach to explore data enhancement by using a right to left model to encourage consensus translations. The existing back-translation method is also one of the data enhancement approaches that leverages monolingual target data to generate synthetic bilingual data. Extending the back-translation approach, our proposed joint-training approach interactively makes data enhancement by boosting source-to-target and target-to-source NMT systems. Eventually, the unified joint training framework, denoted as **ARJT**, is used to integrate the agreement regularization approach, the back translation approach, and the joint training approach to further improve the performance of NMT systems. The evaluation results of the agreement regularization and the unified joint training are listed in the third section of Table 1. Compared to **BT**, our agreement regularization can achieve improvements of 1.34 BLEU points. Adding the joint training can bring this up to a 1.81 gain.

We also explore several variants of our **ARJT** system, denoted as **ARJT2/3/4** in Table 1. We vary the dropout probability in order to explore the interaction between dropout regularization and agreement regularization. Unlike **ARJT**, these variants don't use the validation set for early stopping.

Experimental Results of Data Selection

In addition to our results using the WMT training data, we also explore training our system on a larger corpus. We experimented with 100M parallel sentences drawn from UN data, Open Subtitles and Web crawled data. It is worth noting that the experiments reported in Table 1 were constrained data experiments limited to WMT17 official data only. While the experiments reported in Table 2 are unconstrained systems using additional data.

First we apply word alignment heuristics to filter very noisy data. This filters out around 10% of the data. Then we apply Cross-Entropy data selection [27] and [2] to order the sentences based on their relevance to the CWMT part of the WMT data. We then select a specific number of sentences pairs by rank.

In a separate experiment, we also apply the SentVec similarity filtering, described in Section 3.5, to select the same amount of data and measure its effect. We use a cutoff threshold of the cosine similarity of 0.2. We train the unified bi-lingual encoder on a selected subset of the data that is known to be of good quality and in the relevant domain, specifically, the CWMT data of 9M

sentence pairs. Since the system is trained to translate in both directions, it is effectively trained on on 18M sentence pairs.

Table 2 shows the results of data selection. **Base8K** is using baseline data and back translated data, however it uses a larger model architecture that we found to work better with larger data sets. **Base8K** uses 6-layer transformer with a model size of 1024, a Feed Forward Network size (d_{ff}) of 8192, and 16 heads. All models reported in Table 2 are trained for $300K$ steps with minibatch of 3500 on 8 GPUs. We average the last 20 checkpoints as before and decode with beam size of 8 and length penalty of 1.0 similar to the setup above.

CED1 and **CED2** add 35M sentences and 50M sentences respectively to **Base8k**. **SV1** and **SV2** added the same amount of data selected by SentVec similarity discussed in Section 3.5. **SV3** and **SV4** experimented with varying the dropout ratio to measure its impact with the larger training data and model architecture. Generally the systems using SentVec similarity filtering achieve improvements up to 1.5 BLEU points over **Base8K** and nearly 1 BLEU point as compared to systems using the same amount of CED-selected data. We conclude that SentVec similarity filtering is a helpful approach since it filters out noisy data which is hard to identify. Since SentVec prevents data with partial and low-quality translation from negatively impacting the system. Furthermore, the proposed approach helps select relevant data similar to CWMT data.

SystemID	**Settings**	**BLEU**
Base	Transformer Baseline	24.2
BT	+Back Translation	25.57
Base8K	BT + 8K d_{ff}	26.13
CED1	Base8K + 35M CED + dropout=0.1	26.68
CED2	Base8K + 50M CED + dropout=0.1	26.61
SV1	Base8K + 35M + dropout=0.1	27.60
SV2	Base8K + 50M + dropout=0.1	27.45
SV3	Base8K + 35M + dropout=0.2	27.67
SV4	Base8K + 50M + dropout=0.2	27.49

Table 2: Evaluation Data selection results on the WMT 2017 Chinese-English test set

Experimental Results of Systems Combination

We experiment with system combination of n-best lists generated from various systems discussed above with 8 hypothesis from each system. We use various features to re-rank the systems hypothesis as described in Section 3.6. As shown in Table 3, combining the set of heterogeneous systems are complementary and achieved the highest results. We have experimented with many configurations and features for systems combination, we found out that the most helpful scoring features are: SYS_{Score}, LM_{Score}, $R2L_{score}$, $R2L_{SV}$ and $E2Z_{SV}$. This is quite surprising since the combined systems were focusing on modeling similar features. This may be due to the fact that the models are learning complimentary features, so they have extra capacity for complementing each other.

We think it would be useful to combine all proposed approaches in a single system. However, we leave this as a future work item.

SystemID	**Settings**	**BLEU**
Combo-1	SV1, SV2, SV3	27.84
Combo-2	DLDN2, DLDN3, DLDN4	27.92
Combo-3	ARJT2, ARJT3, ARJT4 + 3 identical systems with different initialization	27.82
Combo-4	SV1, SV2, SV3, ARJT1, ARJT2, ARJT3, DLDN2, DLDN3, DLDN4	28.46
Combo-5	SV1, SV2, SV3, ARJT2, DLDN2, DLDN4	28.32
Combo-6	SV1, SV2, SV4, ARJT2, ARJT3, ARJT4, DLDN2, DLDN3, DLDN4	28.42

Table 3: System combination results on the WMT 2017 Chinese-English test set

#	Ave %	Ave z	System
1	69.0	0.237	COMBO-6
	68.5	0.220	REFERENCE-HT
	68.9	0.216	COMBO-5
	68.6	0.211	COMBO-4
2	67.3	0.141	REFERENCE-PE
3	62.3	-0.094	SOGOU
	62.1	-0.115	REFERENCE-WMT
4	56.0	-0.398	ONLINE-A-1710
	54.1	-0.468	ONLINE-B-1710

Table 4: **Human Evaluation Results** for at least $n \geq 1,827$ assessments per system show that our research systems COMBO-4, COMBO-5, and COMBO-6 achieve human parity according to definition 2 as they are not distinguishable from REFERENCE-HT, which is a human translation. All our research systems significantly outperform REFERENCE-PE, which is based on human post-editing of machine translation output, and the original REFERENCE-WMT, which is again a human translation. # denotes the ranking cluster, *Ave* % the averaged raw score $r \in [0, 100]$, and *Ave* z the standardized z score. $n \geq x$ denotes that we collected at least x assessments per system for the respective evaluation campaign. This is referred to as META-1 in Table 5g.

5 Human Evaluation Results

Table 4 presents the results from our large scale human evaluation campaign. Based on these results we claim that we have achieved human parity according to Definition 2, as our research systems are indistinguishable from human translations.

In the table, systems in higher clusters significantly outperform all systems in lower clusters according to Wilcoxon rank sum test at p-level $p \leq 0.05$, following WMT17. Systems in the same cluster are ordered by z score—which is defined as the signed number of standard deviations an observation is above the mean, computed on the annotator level to address different annotation behavior—but considered tied w.r.t. quality.

5.1 Human Evaluation Setup

As discussed in Section 2 our evaluation methodology is based on *source-based direct assessment* as described in [7]. We use an updated version of Appraise [14], the same tool which is used in the human evaluation campaign for the Conference on Machine Translation (WMT).[11] See [6] for more details on last year's WMT17 results and evaluation.

The main differences to the WMT17 campaign are:

1. Our evaluation is based on quality assessment of translations with respect to the source text, not a reference translation. To do this, we hire bilingual crowd workers;

2. We enforce full system coverage for the evaluation samples. This means that for every segment we get human scores for all systems under investigation;

3. We require redundancy so that for every annotation task (also referred to as "HIT" in other direct assessment publications) we collect scores from three annotators.

The latter two changes have been introduced to strengthen our results, by adding additional redundancy. Direct assessment as an estimator of general system quality does not require these, but in the context of achieving human parity, extra layers of fully comparable segment scores enable more thorough external validation. We intend to release all data related to the final human parity evaluation campaigns, so this data will become available for independent inspection by the research community.

[11] This version of Appraise will also be used to run the WMT18 evaluation campaigns. Source code will be released to the public in time for WMT18, as in previous years.

5.2 Benchmark Translations

We compare our research systems against the following sets of translations. These sets have been kept stable across all evaluation campaigns, allowing us to track research results over time.

Reference-HT vendor-created human translations of *newstest2017*. Translators were instructed to translate from scratch, i.e., without using any online translation engines;[12]

Reference-PE vendor-created human post-editing output, based on Google Translate machine translation results;

Reference-WMT Original *newstest2017* reference released after WMT17. The original WMT17 reference translation for *newstest2017* is known to contain errors, so we decided to add it to the set of evaluated systems. This allows us to get external validation for the quality of our two human references;

Online-A-1710 Microsoft Translator production system, collected on October 16, 2017;

Online-B-1710 Google Translate production system, collected on October 16, 2017;

Sogou The Sogou Knowing NMT system, which performed best at last year's WMT17 Conference on Machine Translation (WMT) shared task on news translation [41].

Note that the benchmark human references were not available to the system developers. Also, the presented set of translation systems affects human-perceived quality (both based on the total number and distribution of quality across systems), so we do not expect scores to be comparable across campaigns. The question of comparability of raw direct assessment scores over time is an open research problem still, so we take a conservative approach and do not compare them. Scores within a single campaign are reliable. We also assume that standardized scores for the same set of translation systems should be fairly comparable.

5.3 Guarding Against Confounds

Whenever trying to draw a conclusion based on a pair of different translations, we must avoid measuring the effects of extraneous variables that can confound the experimental variables we wish to measure [9]. For example, when comparing the translation quality by varying how it is produced (human translation versus automatic translation), we do not wish our measurements of translation quality to be influenced by external factors, e.g., perhaps a human translator did a poor job when translating a few sentences or an automatic translation system happens to be exceptionally good at translating a particular subset of sentences.

In this work, we specifically control for the effects of several potential extraneous variables:

- **Variability of quality measure** *How sensitive is our quality measure (direct assessment) to different subsets of the data?* We answer this by running redundant evaluation campaigns across different subsets of the data.
- **Test set selection** *Would we likely obtain the same result on slightly different test data?* We control for this by running redundant large-scale human evaluation campaigns under several configurations to replicate results (Section 5.4).
- **Annotator errors** *What if some annotators become inattentive, unfairly improving or damaging the score of one system over the other?* To control for this effect, we use rejection sampling when gathering human assessments by occasionally showing annotators examples where one sentence is intentionally and noticeably worse; annotators that fail to detect these are excluded from the data, ensuring that human judgments are high quality.
- **Annotator inconsistency** *What if the annotators produce different scores given the same data? Would using different annotators still lead to the same conclusion?* To control for this, our evaluation campaigns directly include multiple evaluators.

[12] Of course, there are sentences for which the human translation matches Google Translate or Microsoft Translator machine translation output. Relative to the overlap for the post-editing-based reference, this is negligible.

- **Choice of systems** *Was this particular system combination somehow "lucky", or would similar combinations also lead to the same conclusion?* To answer this question, we include multiple system combinations with varying sets of input systems. (Section 5.4)

5.4 Evaluation Campaigns

We conduct the following evaluations:

Annotator variability study To measure this, we repeat the same evaluation campaign three times. All data is collected on the same subset. We allow annotator overlap but do not enforce it. In the end, we had a near complete annotator overlap, likely due to the timing of our campaigns.[13] We refer to this as EVAL ROUND 1, on evaluation sample SUBSET-1;

Data variability study Our data subsets are randomly selected from the source data. Still, the actual subset could affect results in our favor. To counter this, we conduct three additional evaluation campaigns on three completely different subsets of data. We refer to this as EVAL ROUND 2, on evaluation samples SUBSET-2, SUBSET-3, and SUBSET-4.

As the set of systems for all these campaigns does not change, results are theoretically comparable, so we can also report synthesized, joint scores, for both dimensions in isolation and in combined form.

Evaluation campaign parameters are as follows:

- Annotators: 15
- Tasks: 20
- Redundancy: 3
- Tasks per annotator: 4 (about 2 hours of work)
- Systems: 9
- Data points: 4,200 (at least[14] 466 per system)

The set of systems for the final evaluation campaigns consists of the following systems:

- References: REFERENCE-HT, REFERENCE-PE, REFERENCE-WMT
- Production: ONLINE-A-1710, ONLINE-B-1710
- WMT17: SOGOU
- Candidates: COMBO-4, COMBO-5, COMBO-6

After completion of all six evaluation campaigns, we have collected at least 25,200 data points (i.e., segment scores) or at least 2,520 per system. This is comparable to the amount of annotations collected for last year's WMT17 evaluation campaign (2,421 assessments per system). We report results for individual campaigns and our final *synthesized*, joint meta-campaign:

META-1 We combine assessments from evaluation campaigns EVAL ROUND 1a–c, on evaluation sample SubsetB, effectively increasing data points by a factor of 3x. Note that this is fair as result clusters are based on standardized scores which can fairly be computed if all annotators are exposed to exactly the same segments per system.

While it is also possible to combine data across subsets, we choose not to do this as this potentially affects standardization of annotator scores. For META-1, due to the identical assignment of annotators to segments, we have a guarantee that standardization is reliable.

5.5 Annotator Variability Results

SUBSET-1, first iteration Table 5a shows the results of our first evaluation round on SUBSET-1. Note how our research systems outperform SOGOU and both REFERENCE-WMT and REFERENCE-PE. Based on this clustering it becomes clear that there must be quality issues with the original REFERENCE-WMT reference. All three systems COMBO-4, COMBO-5, and COMBO-6 achieve human parity with REFERENCE-HT. We collected at least $n \geq 609$ assessments per system.

[13] To complete so many campaigns in such a short time, it was easier to attract crowd workers when they knew they could earn more by completing several campaigns. Combined with our reliability testing, this motivation likely had a positive impact on annotation fidelity and quality.

[14] Note that as we annotate on unique translation output only, there is a chance that more data points are collected.

Subset-1, second iteration Table 5b shows the results for our second evaluation round on Subset-1. This time, annotators do not see a significant difference between our research systems and Reference-PE. Consequently, Reference-HTand all three systems Combo-4, Combo-5, and Combo-6 end up in the same cluster as Reference-PE. All these systems outperform Sogou and Reference-WMT. As in the previous round, online systems Online-A-1710 and Online-B-1710 perform worst.

Subset-1, third iteration Table 5c shows the results for our third evaluation round on Subset-1. Similar to the second round, we do not observe a significant difference between Reference-PE and our research systems. Again, Reference-HT, all three systems Combo-4, Combo-5, and Combo-6, and Reference-PE end up in the top cluster. Sogou and Reference-WMT end in the third cluster, outperforming Online-A-1710 and Online-B-1710. Again, the latter are not significantly different w.r.t human perceived quality.

5.6 Data Variability Results

Subset-2 Table 5d shows the results for our evaluation on Subset-2. Annotators seem to have a preference for Reference-HT over Combo-4, Combo-5, and Combo-6, but not significantly so. All four systems outperform Reference-PE, which itself outperforms all other systems. Sogou ends up in its own cluster, significantly better than Reference-WMT and the two online systems Online-A-1710 and Online-B-1710. We collected at least $n \geq 607$ assessments per system.

Subset-3 Table 5e shows the results for our evaluation on Subset-3. This one is interesting as it is the only evaluation round which shows Reference-PE on top, based on its z score. Otherwise, we continue to see Reference-HT, Combo-4, Combo-5, and Combo-6 in the top cluster. Sogou and Reference-WMT are indistinguishable for this subset and both outperform the two online systems, Online-A-1710 and Online-B-1710. We collected at least $n \geq 610$ assessments per system.

Subset-4 Table 5f shows the results for our evaluation on Subset-4. Again, our research systems Combo-4, Combo-5, and Combo-6 are indistinguishable from Reference-HT and Reference-PE. There is no significant difference in quality between these five systems. Sogou and Reference-WMT outperform the online systems Online-A-1710 and Online-B-1710. We collected at least $n \geq 649$ assessments per system.

#	Ave %	Ave z	System
1	69.9	0.256	Combo-6
	69.8	0.233	Combo-4
	69.9	0.230	Combo-5
	68.6	0.186	Reference-HT
	67.6	0.129	Reference-PE
2	63.3	-0.095	Sogou
	62.1	-0.132	Reference-WMT
3	57.0	-0.383	Online-A-1710
	54.1	-0.494	Online-B-1710

(a) Subset-1, $n \geq 609$

#	Ave %	Ave z	System
1	68.6	0.233	Reference-HT
	68.6	0.225	Combo-6
	68.6	0.217	Combo-5
	68.3	0.207	Combo-4
	67.4	0.154	Reference-PE
2	61.9	-0.105	Sogou
	62.1	-0.113	Reference-WMT
3	55.7	-0.399	Online-A-1710
	53.9	-0.468	Online-B-1710

(b) Subset-1, second iteration

#	Ave %	Ave z	System
1	68.5	0.240	Reference-HT
	68.4	0.229	Combo-6
	68.1	0.201	Combo-5
	67.7	0.194	Combo-4
	66.8	0.141	Reference-PE
2	61.8	-0.083	Sogou
	62.0	-0.100	Reference-WMT
3	55.2	-0.413	Online-A-1710
	54.3	-0.442	Online-B-1710

(c) Subset-1, third iteration

#	Ave %	Ave z	System
1	68.6	0.212	Reference-HT
	68.2	0.200	Combo-5
	67.9	0.182	Combo-4
	67.9	0.177	Combo-6
2	64.8	0.044	Reference-PE
	62.5	-0.061	Sogou
3	59.6	-0.200	Reference-WMT
	58.4	-0.277	Online-A-1710
	55.7	-0.353	Online-B-1710

(d) Subset-2, $n \geq 607$

#	Ave %	Ave z	System
1	67.4	0.251	Reference-HT
	67.1	0.247	Reference-PE
	65.3	0.147	Combo-6
	64.9	0.106	Combo-4
	64.3	0.091	Combo-5
2	61.1	-0.065	Sogou
	59.6	-0.119	Reference-WMT
3	55.3	-0.351	Online-A-1710
	54.4	-0.377	Online-B-1710

(e) Subset-3, $n \geq 650$

#	Ave %	Ave z	System
1	66.6	0.254	Reference-HT
	65.2	0.179	Combo-6
	64.4	0.151	Combo-5
	64.2	0.147	Combo-4
	63.4	0.127	Reference-PE
2	60.5	-0.030	Sogou
	60.1	-0.074	Reference-WMT
3	53.4	-0.367	Online-A-1710
	51.7	-0.455	Online-B-1710

(f) Subset-4, $n \geq 649$

#	Ave %	Ave z	System
1	69.0	0.237	Combo-6
	68.5	0.220	Reference-HT
	68.9	0.216	Combo-5
	68.6	0.211	Combo-4
2	67.3	0.141	Reference-PE
3	62.3	-0.094	Sogou
	62.1	-0.115	Reference-WMT
4	56.0	-0.398	Online-A-1710
	54.1	-0.468	Online-B-1710

(g) Meta-1, $n \geq 1,827$

Table 5: Complete results for our three iterations over Subset-1 (5a, 5b, 5c) and our evaluation campaigns for Subset-2 (5d), Subset-3 (5e), and Subset-4 (5f). We also show results for combined data for Meta-1 (5g) combining annotations from all iterations over Subset-1. # denotes the ranking cluster, *Ave* % the averaged raw score $r \in [0, 100]$, and *Ave* z the standardized z score. $n \geq x$ denotes that we collected at least x assessments per system for the respective evaluation campaign. All campaigns involved $a = 15$ annotators. Systems in higher clusters significantly outperform all systems in lower clusters according to Wilcoxon rank sum test at p-level $p \leq 0.05$, following WMT17. Systems in the same cluster are ordered by z score but considered tied w.r.t. quality.

System	refs=1			refs=2			refs=3
	WMT	PE	HT	WMT+PE	WMT+HT	PE+HT	WMT+PE+HT
Online-A-1710	24.38	28.82	17.12	36.53	32.17	35.33	41.21
Online-B-1710	33.56	46.97	17.70	56.45	40.55	51.78	59.37
Sogou	26.37	30.69	19.71	38.67	35.47	38.19	44.18
Combo-4	28.30	29.79	20.47	39.53	37.73	38.43	45.62
Combo-5	28.18	29.61	20.48	39.32	37.54	38.15	45.32
Combo-6	28.07	29.90	20.70	39.39	37.77	38.45	45.64

Table 6: BLEU scores against single or multiple references. WMT is Reference-WMT, PE is Reference-PE, HT is Reference-HT. Scoring based on sacreBLEU v1.2.3, with signature `BLEU+case.mixed+numrefs.1+smooth.exp+tok.13a+version.1.2.3` for refs=1. Signature changes to `numrefs.2` and `numrefs.3` for refs=2 and refs=3, respectively. Note how different scores for Reference-WMT and Reference-PE are compared to Reference-HT and how these compare to our findings reported in Table 5. This emphasizes the need for human evaluation.

5.7 Data Release

We have released[15] all data from the human evaluation campaigns to 1) allow external validation of our claim of having achieved human parity and 2) to foster future research by releasing two additional human references for the Reference-WMT test set.

The release package contains the following items:

New references for newstest2017 Two new references for newstest2017, one based on human translation from scratch (Reference-HT), the other based on human post-editing (Reference-PE). Table 6 reports the BLEU scores for single and multi reference use with sacreBLEU;

Human parity translations Output generated by our research systems Combo-4, Combo-5, and Combo-6;

Online translations Output from online machine translation service Online-A-1710, collected on October 16, 2017;

Human evaluation data All data points collected in our human evaluation campaigns. This includes annotations for Subset-1, Subset-2, Subset-3, and Subset-4. We share the (anonymized) annotator IDs, segment IDs, system IDs, type ID (either `TGT` or `CHK`, the second being a repeated judgment for the first), raw scores $r \in [0, 100]$, as well as annotation start and end times.

We do not redistribute the following items:

Reference-WMT test data This is publicly available from the WMT17 website[16]. In this work, we used the source `newstest2017-zhen-src.zh` and the reference (as Reference-WMT) `newstest2017-zhen-ref.en`;

Sogou translation This is publicly available from the WMT17 website as well[17]. We used `newstest2017.SogouKnowing-nmt.5171.zh-en` (as Sogou).

The Appraise repository on GitHub[18] contains code to recompute result clusters. We share this data in the hope that the research community might find it useful and also to ensure greatest possible transparency regarding the generation of the results presented in this paper.

[15] All Translator human parity data is available here: http://aka.ms/Translator-HumanParityData
[16] http://data.statmt.org/wmt17/translation-task/test.tgz
[17] http://data.statmt.org/wmt17/translation-task/wmt17-submitted-data-v1.0.tgz
[18] https://github.com/cfedermann/Appraise

6 Human Analysis

Lastly, a preliminary human error analysis was conducted over the output of the COMBO-6 system (the system that achieved the best results). We randomly sampled 500 sentences and annotated each translation with whether a specific error type was present. Following [40], we use 9 categories: Missing Words, Word Repetition, Named Entity, Word Order, Incorrect Words, Unknown Words, Collocation, Factoid, and Ungrammatical. The Named-Entity category is further subdivided into Person, Location, Organization, Event, and Other.

Error Category	**Fraction [%]**
Incorrect Words	7.64
Ungrammatical	6.33
Missing Words	5.46
Named Entity	4.38
Person	1.53
Location	1.53
Organization	0.66
Event	0.22
Other	0.44
Word Order	0.87
Factoid	0.66
Word Repetition	0.22
Collocation	0.22
Unknown Words	0

Table 7: Error distribution, as fraction of sentences that contain specific error categories.

Table 7 shows the distribution of the annotated errors as the fraction of sentences containing a specific error category. The four major error types are Missing words, Incorrect Words, Ungrammatical, and Named Entity. Each accounts for roughly 5% of errors. This indicates that there is still room to improve machine translation quality via various approaches, such as modeling Missing Words [36, 15], integration of high quality data for named-entity translation, as well as domain and topic adaptation for the issues of incorrect words and ungrammaticality.

7 Discussion and Future Work

In this paper, we described the techniques used in the latest Microsoft machine translation system to reach a new state-of-the-art. Our evaluation found that our system has reached parity with professional human translations on the WMT 2017 Chinese to English news task, and exceeds the quality of crowd-sourced references.

We exploited the dual nature of the translation problem to better utilize parallel data as well as monolingual data in a more principled way. We utilized joint training of source-to-target, and target-to-source systems to further improve on the duality of the translation task. We addressed the exposure bias problem in two ways: by two-pass decoding using Deliberation networks, as well as by agreement regularization and joint training of left-to-right, right-to-left systems. We trained a bilingual encoder to obtain bilingual sentence representations used to filter noisy data and select relevant data. We also found significant gains from combining multiple heterogeneous systems.

We addressed the problem of defining and measuring the quality of human translations and near-human machine translations. We found that as translation quality has dramatically improved, automatic reference-based evaluation metrics have become increasingly problematic. We used direct human annotation to measure the quality of both human and machine translations.

We wish to acknowledge the tremendous progress in sequence-to-sequence modeling made by the entire research community that paved the road for this achievement. We have introduced a few new approaches that helped us to reach human parity for WMT2017 Chinese to English news translation task. At the same time, much work remains to be done, especially in domains and language-pairs that do not benefit from huge amounts of available data.

References

[1] Artetxe, M., Labaka, G., Agirre, E., and Cho, K. Unsupervised neural machine translation. In *International Conference on Learning Representations* (2018).

[2] Axelrod, A., He, X., and Gao, J. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Stroudsburg, PA, USA, 2011), EMNLP '11, Association for Computational Linguistics, pp. 355–362.

[3] Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Belinkov, Y., and Bisk, Y. Synthetic and natural noise both break neural machine translation. *CoRR abs/1711.02173* (2017).

[5] Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS* (2015), pp. 1171–1179.

[6] Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huang, S., Huck, M., Koehn, P., Liu, Q., Logacheva, V., Monz, C., Negri, M., Post, M., Rubino, R., Specia, L., and Turchi, M. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers* (Copenhagen, Denmark, September 2017), Association for Computational Linguistics, pp. 169–214.

[7] Cettolo, M., Federico, M., Bentivogli, L., Niehues, J., Stüker, S., Sudoh, K., Yoshino, K., and Federmann, C. Overview of the iwslt 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT)* (Tokyo, Japan, December 2017), IWSLT, pp. 2–12.

[8] Cherry, C., and 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* (Stroudsburg, PA, USA, 2012), NAACL HLT '12, Association for Computational Linguistics, pp. 427–436.

[9] Clark, J. H., Dyer, C., Lavie, A., and Smith, N. A. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2* (2011), Association for Computational Linguistics, pp. 176–181.

[10] Denkowski, M., and Lavie, A. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation* (2011).

[11] Devlin, J. Sharp models on dull hardware: Fast and accurate neural machine translation decoding on the cpu. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Copenhagen, Denmark, September 2017), Association for Computational Linguistics, pp. 2810–2815.

[12] Dreyer, M., and Marcu, D. HyTER: Meaning-Equivalent Semantics for Translation Evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2012), Association for Computational Linguistics, pp. 162–171.

[13] Edgington, E. S. Validity of Randomization Tests for One-subject Experiments. *Journal of Educational Statistics 5*, 3 (1980), 235–251.

[14] Federmann, C. Appraise: An open-source toolkit for manual evaluation of machine translation output. *The Prague Bulletin of Mathematical Linguistics 98* (September 2012), 25–35.

[15] Feng, S., Liu, S., Yang, N., Li, M., Zhou, M., and Zhu, K. Q. Improving attention modeling with implicit distortion and fertility for machine translation. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan* (2016), pp. 3082–3092.

[16] Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122* (2017).

[17] Gu, J., Hassan, H., Devlin, J., and Li, V. Universal neural machine translation for extremely low resource languages.

[18] Hassan, H., Elaraby, M., and Tawfik, A. Y. Synthetic data for neural machine translation of spoken-dialects.

[19] He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., and Ma, W.-Y. Dual learning for machine translation. In *Advances in Neural Information Processing Systems* (2016), pp. 820–828.

[20] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385* (2015).

[21] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* (2016).

[22] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014).

[23] Lample, G., Conneau, A., Denoyer, L., and Ranzato, M. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations* (2018).

[24] Lin, J., Xia, Y., Qin, T., Chen, Z., and Liu, T.-Y. Conditional image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2018).

[25] Luo, P., Wang, G., Lin, L., and Wang, X. Deep dual learning for semantic image segmentation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017* (2017), pp. 2737–2745.

[26] Mann, H. B., and Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[27] Moore, R. C., and Lewis, W. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers* (Stroudsburg, PA, USA, 2010), ACLShort '10, Association for Computational Linguistics, pp. 220–224.

[28] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (2002), Association for Computational Linguistics, pp. 311–318.

[29] Sennrich, R., Haddow, B., and Birch, A. Improving neural machine translation models with monolingual data. *arXiv preprint arXiv:1511.06709* (2015).

[30] Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909* (2015).

[31] Shen, W., and Liu, R. Learning residual images for face attribute manipulation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), IEEE, pp. 1225–1233.

[32] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Driessche, G. V. D., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., and et al. Mastering the game of go with deep neural networks and tree search. *Nature, vol. 529, pp. 484489* (2016).

[33] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas* (2006), vol. 200.

[34] Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *NIPS* (2014).

[35] Tang, D., Duan, N., Qin, T., and Zhou, M. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027* (2017).

[36] Tu, Z., Lu, Z., Liu, Y., Liu, X., and Li, H. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811* (2016).

[37] van der Wees, M., Bisazza, A., and Monz, C. Dynamic data selection for neural machine translation. *CoRR abs/1708.00712* (2017).

[38] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).

[39] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NIPS* (2017).

[40] Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. Error analysis of statistical machine translation output. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.* (2006), pp. 697–702.

[41] Wang, Y., Cheng, S., Jiang, L., Yang, J., Chen, W., Li, M., Shi, L., Wang, Y., and Yang, H. Sogou neural machine translation systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017* (2017), pp. 410–415.

[42] Wang, Y., Li, X., Cheng, S., Jiang, L., Yang, J., Chen, W., Shi, L., Wang, Y., and Yang, H. Sogou neural machine translation systems for wmt17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers* (Copenhagen, Denmark, September 2017), Association for Computational Linguistics, pp. 410–415.

[43] Wang, Y., Xia, Y., Zhao, L., Bian, J., Qin, T., Liu, G., and Liu, T. Dual transfer learning for neural machine translation with marginal distribution regularization. In *AAAI* (2018).

[44] Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin 1*, 6 (1945), 80–83.

[45] Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv e-prints* (Sept. 2016).

[46] Xia, Y., Bian, J., Qin, T., Yu, N., and Liu, T.-Y. Dual inference for machine learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17* (2017), pp. 3112–3118.

[47] Xia, Y., Qin, T., Chen, W., Bian, J., Yu, N., and Liu, T. Dual supervised learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), pp. 3789–3798.

[48] Xia, Y., Tian, F., Wu, L., Lin, J., Qin, T., Yu, N., and Liu, T.-Y. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems* (2017), pp. 1782–1792.

[49] Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D., and Zweig, G. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 25*, 12 (2017), 2410–2423.

[50] Yi, Z., Zhang, H., Tan, P., and Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. *ICCV* (2017).

[51] Zhang, Z., Liu, S., Li, M., Zhou, M., and Chen, E. Joint training for neural machine translation models with monolingual data, 2018.

[52] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593* (2017).

[53] Zoph, B., Yuret, D., May, J., and Knight, K. Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201* (2016).