Principled Paraphrase Generation with Parallel Corpora

Aitor Ormazabal[1], **Mikel Artetxe**[2], **Aitor Soroa**[1], **Gorka Labaka**[1], **Eneko Agirre**[1]
[1]HiTZ Center, University of the Basque Country (UPV/EHU)
[2]Meta AI
{aitor.ormazabal,a.soroa,gorka.labaka,e.agirre}@ehu.eus
artetxe@fb.com

Abstract

Round-trip Machine Translation (MT) is a popular choice for paraphrase generation, which leverages readily available parallel corpora for supervision. In this paper, we formalize the implicit similarity function induced by this approach, and show that it is susceptible to non-paraphrase pairs sharing a single ambiguous translation. Based on these insights, we design an alternative similarity metric that mitigates this issue by requiring the entire translation distribution to match, and implement a relaxation of it through the Information Bottleneck method. Our approach incorporates an adversarial term into MT training in order to learn representations that encode as much information about the reference translation as possible, while keeping as little information about the input as possible. Paraphrases can be generated by decoding back to the source from this representation, without having to generate pivot translations. In addition to being more principled and efficient than round-trip MT, our approach offers an adjustable parameter to control the fidelity-diversity trade-off, and obtains better results in our experiments.

1 Introduction

Paraphrase generation aims to generate alternative surface forms expressing the same semantic content as the original text (Madnani and Dorr, 2010), with applications in language understanding and data augmentation (Zhou and Bhat, 2021). One popular approach is to use an MT system to translate the input text into a pivot language and back (Wieting and Gimpel, 2018; Mallinson et al., 2017; Roy and Grangier, 2019). While it intuitively makes sense that translating to another language and back should keep the meaning of a sentence intact while changing its surface form, it is not clear what exactly would be considered a paraphrase by such a system.

In this work, we show that the probability of a paraphrase x_p given a source sentence x_s under a round-trip MT system can be naturally decomposed as $P(x_p|x_s) = P(x_p)S(x_p, x_s)$, where S is a symmetric similarity metric over the paraphrase space and $P(x_p)$ the probability of x_p. We argue that this similarity function is not appropriate in the general case, as it can assign a high score to sentence pairs that share an ambiguous translation despite not being paraphrases of each other. This phenomenon is illustrated in Figure 2, where x_s and x_p share a confounding translation without gender marker.

So as to address this issue, we design an alternative similarity function that requires the entire translation distribution to match, and develop a relaxation of it through the Information Bottleneck (IB) method. We implement this approach using an adversarial learning system depicted in Figure 1.

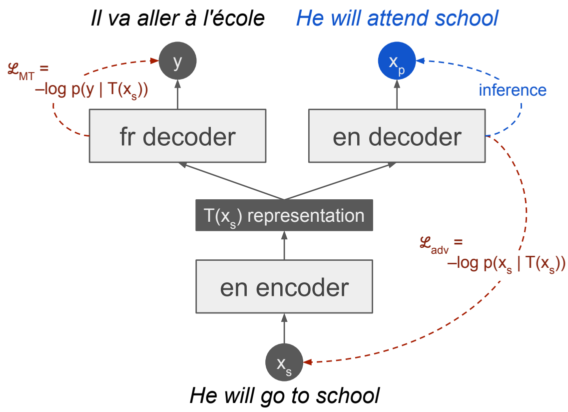

Figure 1: **Proposed system.** Given the input x_s, we aim to learn a representation $T(x_s)$ that encodes as much information as possible about it's reference translation y (ensuring that the meaning is preserved), and as little information as possible about x_s itself (ensuring that surface information is removed). We achieve this through adversarial learning, where the encoder minimizes $\lambda\mathcal{L}_{MT} - (1-\lambda)\mathcal{L}_{adv}$ and the two decoders minimize $\mathcal{L}_{MT}$ and $\mathcal{L}_{adv}$. At inference time, we couple the English encoder and decoder to generate a paraphrase x_p which, being conditioned on $T(x)$, will preserve the meaning of x_s but use a different surface form.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 1621 - 1638

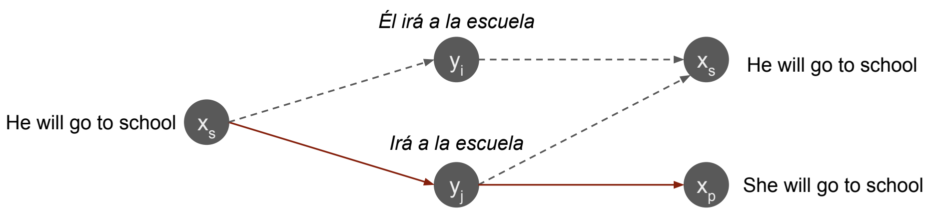

Figure 2: **Confounding translation problem in round-trip MT.** *"Irá a la escuela"* does not mark the gender of the subject due to ellipsis, and it is thus a valid translation of both *"He will go to school"* and *"She will go to school"*. As a consequence, round-trip MT could generate *"She will go to school"* as a paraphrase of *"He will go to school"*. Our approach mitigates this issue by requiring the full translation distribution to match.

Our model combines an encoder that, for a given sentence, removes the information that is not relevant to predict its translation, and a decoder that reconstructs a paraphrase from this encoding. In addition to being more principled, our approach is more efficient than round-trip MT at inference, can be tuned to favor fidelity or diversity, and achieves a better trade-off between the two. Our code is freely available [1].

2 Related Work

We next review the paraphrase generation literature (§2.1), and describe the information bottleneck method (§2.2), which is the basis of our proposal.

2.1 Paraphrase generation

Early work on paraphrasing focused on retrieval methods, either extracting plausible sentences from large corpora for generation (Barzilay and McKeown, 2001; Bannard and Callison-Burch, 2005), or identifying paraphrase pairs from weakly aligned corpora to create paraphrase datasets (Coster and Kauchak, 2011; Dolan et al., 2004). More recently, neural approaches for paraphrase generation have dominated the field. We classify these methods according to the type of supervision they use.

Monolingual corpora. These systems are trained in an unsupervised fashion using unlabeled monolingual corpora. They usually employ an information bottleneck, with the goal of encoding semantic information in the latent space. Approaches include Variational Autoencoders (VAE) (Bowman et al., 2016), VAEs with Vector Quantization (Roy and Grangier, 2019), and latent bag-of-words models (Fu et al., 2019). Huang and Chang (2021) disentangle semantic and syntactic content in the latent space through a bag of words representation, which allows for syntactically controllable generation.

Parallel corpora. These systems are trained on pairs of parallel sentences in two languages. Most of these methods are based on round-trip MT, where a sentence is translated to a pivot language and back in order to obtain a paraphrase. Hu et al. (2019) add lexical constraints to the MT decoding procedure to obtain better paraphrases. Mallinson et al. (2017) generate not one but multiple pivot sentences and use a fusion-in-decoder strategy.

Paraphrase corpora. These systems are trained in a supervised manner over pairs or clusters of paraphrases. When such data is available, training a regular sequence-to-sequence model is a strong baseline (Egonmwan and Chali, 2019). Kumar et al. (2019) add submodular optimization to improve paraphrase diversity. Some VAE-based methods also leverage paraphrase clusters to learn a latent representation that disentangles meaning and form (Iyyer et al., 2018; Kumar et al., 2020; Hosking and Lapata, 2021; Chen et al., 2019). Most of these methods require a syntactic exemplar for generation, and assume that all surface forms are valid for all sentences. Hosking and Lapata (2021) do away with this assumption in the context of question paraphrasing, predicting a valid syntactic embedding from a discrete set at test time.

While it is paraphrase corpora that offers the strongest supervision, such data is hard to obtain and usually restricted to narrow domains like Quora Question Pairs, WikiAnswers and Twitter (Hosking and Lapata, 2021; Kumar et al., 2019; Egonmwan and Chali, 2019). In contrast, parallel corpora is widely available, while offering a stronger training signal than monolingual corpora. For that reason, round-trip MT is a common choice when paraphrases are needed for downstream tasks (Xie et al.,

[1] https://github.com/aitorormazabal/paraphrasing-from-parallel

2020; Artetxe et al., 2020), as well as a common baseline in the paraphrasing literature (Hosking and Lapata, 2021; Roy and Grangier, 2019).[2] Our work focuses on this class of systems, identifying the limitations of round-trip MT and proposing a more principled alternative.

2.2 The Information Bottleneck Method

Given two random variables X, Y, the Information Bottleneck (IB) method (Tishby et al., 1999) seeks to learn a representation $T(X)$ that minimizes the Mutual Information (MI) between T and X, while preserving a minimum MI between T and Y. That is, the objective $I(X,T)\ s.t.\ I(T,Y) \geq \gamma$ is minimized. Since the MI is usually impossible to calculate exactly for neural representations, a common approach is to use variational methods, that turn the estimation problem into an optimization one. This can be done by adding a neural decoder on top of the representation, and training the entire system end-to-end (Poole et al., 2019). This is the approach we follow in this work.

3 Characterizing Round-trip MT

Let X be a random variable representing a sequence in the source language, and Y be a random variable representing its translation into a pivot language.[3] Given an input sequence $x_s \in X$, we can use round-trip MT to generate a paraphrase $x_p \in X$ by translating x_s into the pivot language and back, according to the forward and backward translation models $P(y|x_s)$ and $P(x_p|y)$. As such, we can formulate the probability of round-trip MT generating a particular paraphrase x_p by marginalizing over the set of possible pivot translations:

$$P(x_p|x_s) = \sum_{y \in Y} P(y|x_s)P(x_p|y) \tag{1}$$

In what follows, we will characterize the paraphrases produced by this approach, i.e. the properties that x_p needs to meet in relation to x_s for $P(x_p|x_s)$ to be high.[4]

[2] Round-trip MT has also been used to generate synthetic paraphrase corpora (Wieting and Gimpel, 2018).

[3] For convenience, we will also use X and Y to refer to the set of source and target language sequences, and abbreviate probabilities of the form $P(X = x)$ as $P(x)$.

[4] Some round-trip MT systems do not consider all possible translations into the pivot language, but only a subset of them (Mallinson et al., 2017). In that case, the sum in Eq. 1 goes over $y \in \{y_1, ..., y_K\}$, and we need to introduce a partition $Z = \sum_{y \in \{y_1,...,y_K\}} P(y|x_s)$ to normalize the probabilities. However, the fundamental analysis in this section still applies. Refer to Appendix A for more details.

By applying Bayes' rule, we can rewrite Eq. 1 as follows:

$$P(x_p|x_s) = P(x_p) \underbrace{\sum_{y \in Y} \frac{P(y|x_s)P(y|x_p)}{P(y)}}_{S_{MT}(x_p,x_s)} \tag{2}$$

The sum on the right hand side can be interpreted as a symmetric similarity function, $S_{MT}(x_p, x_s) = S_{MT}(x_s, x_p) = \sum_y \frac{P(y|x_s)P(y|x_p)}{P(y)}$, which measures the likelihood of two sentences to be actual paraphrases. The probability of x_p given x_s then becomes $P(x_p|x_s) = P(x_p)S_{MT}(x_p, x_s)$, which is the similarity between x_s and x_p, weighted by the marginal probability of x_p.

But when are x_s and x_p considered similar according to the above definition of $S_{MT}(x_s, x_p)$? Intuitively, S_{MT} is a measure of the *overlap* between the conditional distributions that x_s and x_p induce over Y. This will be highest when $P(y|x_s)P(y|x_p)$ is as large as possible for as many y as possible. At the same time, $P(y|x_s)P(y|x_p)$ will be high when both $P(y|x_s)$ and $P(y|x_p)$ are high, that is, when y is a probable translation of both x_s and x_p. This captures the intuition that two sentences are similar when they can be translated into the same text in the pivot language.

But what if x_s and x_p have one particular high-probability translation y_j in common, but differ in the rest? As illustrated in Figure 2, this can happen when y_j is ambiguous in the target language and can mean both x_s and x_p, even if x_s and x_p are not equivalent (e.g., when x_s uses the masculine form, x_p the feminine form, and y_j does not mark the gender). In this case, the sum $\sum_y \frac{P(y|x_s)P(y|x_p)}{P(y)}$ will be dominated by $\frac{P(y_j|x_s)P(y_j|x_p)}{P(y_j)}$, which will be high when both $P(y_j|x_s)$ and $P(y_j|x_p)$ are high.

We can thus conclude that the implicit similarity function underlying round-trip MT is flawed, as it assigns a high score to a pair of sequences (x_s, x_p) that have an ambiguous translation in common. As a consequence, round-trip MT will generate x_p as a paraphrase of x_s with a high probability, even if the two sequences have a different meaning.

4 Principled Paraphrasing

As shown in the previous section, the implicit similarity function induced by round-trip MT is not adequate in the general case, as it assigns a high score to pairs of sequences that share a single translation, despite differing in the rest. So as to address this, we can define an alternative similarity function that requires the entire translation distribution to match:

$$S(x_p, x_s) = \begin{cases} 1 & P(y|x_p) = P(y|x_s) \forall y \in Y \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and use it to replace S_{MT} in Eq. 2 so that $P(x_p|x_s) \propto P(x_p)S(x_p, x_s)$. However, this definition is too strict, as it is virtually impossible that $P(y|x_p)$ and $P(y|x_s)$ are exactly the same for all $y \in Y$.[5] In 4.1, we define a relaxation of it through the IB method, which introduces an adjustable parameter β to control how much we deviate from it. In 4.2, we characterize the paraphrases generated by this approach, showing that they are less susceptible to the problem of confounding translations described in the previous section.

4.1 IB-based relaxation

So as to implement the similarity function in Eq. 3, we will use the IB method to learn an encoding T for X such that the following holds:

$$S(x_p, x_s) = \frac{P(x_p|T(x_s))}{P(x_p)Z(x_s)} \tag{4}$$

where $Z(x_s)$ is a normalizer that does not depend on the paraphrase candidate x_p.

As seen in §2.2, given a source variable X and a target variable Y, the IB method seeks to find an encoding $T(X)$ that minimizes the MI with X (maximizing compression), while preserving a certain amount of information about Y:

$$\min_T I(X, T) \ s.t \ I(T, Y) \geq \gamma. \tag{5}$$

This constrained minimization is achieved by introducing a Lagrange multiplier β and minimizing

$$\min_T I(X, T) - \beta I(T, Y). \tag{6}$$

As $\beta \to \infty$, all the information about Y is preserved and the IB method learns a minimal sufficient statistic T, that is, an encoding that satisfies $I(T, Y) = I(X, Y)$ while achieving the lowest $I(T, X)$ possible. The following theorem states that such a minimal sufficient statistic T induces the similarity function in Eq. 3 (proof in Appendix C):

Theorem 1. *Suppose the random variable X represents a sentence in the source language, Y represents its translation, and T is a minimal sufficient statistic of X with respect to Y. Let x_p and x_s be a pair of sentences in the source language. Then, $P(x_p|T(x_s)) = P(x_p)\frac{S(x_p,x_s)}{Z(x_s)}$, where S is given by Equation 3, and Z is a normalizing factor that does not depend on x_p.*

Thus, as $\beta \to \infty$ the IB method approximates the similarity metric S. In practice, when β is set to a fixed finite number, losing some information about the target variable is allowed, and a relaxation of the metric S is learned instead.

4.2 Characterizing IB-based paraphrasing

We will next analyze the relaxation of S induced by the IB method. We will characterize what kind of sentences are considered paraphrases by it, showing that it is less susceptible to the problem of confounding translations found in round-trip MT (§3). Derivations for the results in this section, as well as alternative bounds and broader discussion can be found in Appendix B.

As seen in §4.1, we define paraphrase probabilities given an encoding T as $P(x_p|T(x_s)) = P(X = x_p|T(X) = T(x_s))$, which can only be non-zero if $T(x_p) = T(x_s)$. This means that the encoding T will partition the source space into a collection of paraphrase clusters according to its value. Mathematically, given the equivalence relation $x_1 \sim x_2 \iff T(x_1) = T(x_2)$, only sentence pairs within the same equivalence class will have non-zero paraphrase probabilities. We then have the following theorem:

Theorem 2. *Suppose T is a solution of the IB optimization problem $\min_T I(X, T) \ s.t \ I(T, Y) \geq \gamma$, and $\epsilon = I(X, Y) - \gamma$. If $\mathcal{A}$ is the partition on X induced by T, we have:*

$$\sum_{A \in \mathcal{A}} \max_{x_1, x_2 \in A} \frac{P(x_1)P(x_2)}{2(P(x_1) + P(x_2))} \cdot D_1(P_{Y|x_1}, P_{Y|x_2})^2 \leq \epsilon, \tag{7}$$

where D_1 is the L_1 norm distance.

It is easy to see that, when $\epsilon = 0$, corresponding to $\gamma = I(X, Y)$ and $\beta \to \infty$, this forces all distances to be zero. In that case, only sentences with identical translation distributions are considered paraphrases, in accordance with Theorem 1.

In the general case, Theorem 2 states that the L_1 distance between the translation distributions of

[5] One reason is that we use empirical estimates of $P(y|x_p)$ and $P(y|x_s)$, which will deviate from the ground truth.

sentences that are considered paraphrases cannot be high, as it will be bounded by a function of ϵ. While the S_{MT} metric in §3 can be dominated by a high-probability term and effectively ignore differences in probability for the less likely translations, the L_1 norm gives equal importance to differences in probability for every translation candidate. Thanks to this, the resulting system will be less susceptible to the problem of confounding translations.

5 Proposed System

In this section, we describe a practical implementation of the IB-based paraphrasing approach defined theoretically in §4.

As illustrated in Figure 1, our system can be seen as an extension of a regular encoder-decoder MT architecture with an additional adversarial decoder, which is trained with an auto-encoding objective to reconstruct the original sentence x_s from the encoder representation $T(x_s)$. The encoder is trained to minimize the cross-entropy loss of the MT decoder, while maximizing the loss of the adversarial decoder. This way, the encoder is encouraged to remove as much information about x_s as possible, while retaining the information that is necessary to predict its reference translation y. Thanks to this, $T(x_s)$ should capture the semantic content of x_s (which is relevant to predict y), without storing additional surface information (which is not relevant to predict y). Once the model is trained, the adversarial decoder can be used to generate paraphrases of x_s from this representation $T(x_s)$.

This adversarial architecture can be interpreted as an implementation of the IB method as follows. Following Poole et al. (2019), we start by adding a decoder $q(y|t)$ on top of the encoder $T(x)$, and rewrite the $I(T,Y)$ term as:

$$\begin{aligned} I(T,Y) &= \mathbb{E}_{P(y,t)}\left[\log \frac{q(y|t)}{P(y)}\right] \\ &\quad + \mathbb{E}_{P(t)}[KL(P(y|t)||q(y|t))] \\ &\geq \mathbb{E}_{P(y,t)}\left[\log q(y|t)\right] + h(Y), \end{aligned} \tag{8}$$

where equality will hold if q is the true conditional distribution $q(y|t) = P(y|t)$, and h is the differential entropy. If we parametrize T and q by a neural network encoder-decoder architecture the $\mathbb{E}_{P(y,t)}\left[\log q(y|t)\right]$ term in Eq. 8, can be rewritten as $\mathbb{E}_{P(y,x)}\left[\log q(y|T(x))\right]$, which is precisely the log likelihood of the data distribution of X, Y given by P. In other words, by training the encoder-decoder to maximize Eq. 8, we are implicitly maximizing the mutual information $I(T,Y)$.

Similarly, one can approximate

$$\begin{aligned} I(X,T) &\geq \mathbb{E}_{P(x,t)}\left[\log q(x|t)\right] + h(X) \\ &= \mathbb{E}_{P(x)}\left[\log q(x|T(x))\right] + h(X), \end{aligned} \tag{9}$$

where equality will hold when q is the true conditional distribution and $q(x|T(x)) = P(x|T(x))$. Thus, given an ideal decoder q that perfectly approximates the conditional distributions $q(x|T(x))$ and $q(y|T(x))$, the IB minimization problem is equivalent to minimizing

$$\begin{aligned} &\mathbb{E}_{p(x)}\left[\log q(x|T(x))\right] - \beta \mathbb{E}_{P(y,t)}\left[\log q(y|t)\right] \\ &= \mathbb{E}_{P(x,y)}\left[\log q(x|T(x)) - \beta \log q(y|T(x))\right]. \end{aligned} \tag{10}$$

In practice, we parametrize both the encoder T and the decoder q with transformer neural networks, and learn them from a parallel corpus. Since $\log q(y|T(x))$ is a lower bound of $I(T,Y) - h(Y)$, maximizing this term is theoretically sound. Minimizing $\mathbb{E}_{P(x)}\left[\log q(x|T(x))\right]$, on the other hand, amounts to minimizing a lower bound, which, while not as theoretically solid, is common practice in the variational optimization literature (Chen et al., 2018; Kim and Mnih, 2018).

Finally, we reparametrize Eq. 10 by setting $\lambda = \frac{\beta}{1+\beta}$ to obtain the equivalent minimization objective

$$\begin{aligned} \mathcal{L}(T,q) = \mathbb{E}_{P(x,y)}[-\lambda \log q(y|T(x)) \\ +(1-\lambda)\log q(x|T(x))] = \\ \lambda \mathcal{L}_{MT}(T,q) - (1-\lambda)\mathcal{L}_{Adv}(T), \end{aligned} \tag{11}$$

where $\mathcal{L}_{MT}$ is the regular MT loss of cross-entropy with the translation target, and $\mathcal{L}_{Adv}$ is the cross-entropy with the source sentence (see Figure 1).[6] We thus observe that the proposed adversarial training architecture approximates the IB method. The

[6]We make the adversarial term a function of T only in the minimization objective, as the gradient from the adversarial term is only propagated to the encoder. The adversarial decoder is independently trained to predict the source from the encoded representation.

setting $\beta \to \infty$ corresponds to $\lambda \to 1$, where the optimal solution is a minimal sufficient statistic.

During training, the expectation in Eq. 11 is approximated by sampling batches from the training data. Care must be taken when optimizing the loss, as we do not want to propagate gradients of the adversarial loss to the adversarial decoder. If we did, a trivial way to minimize $(1-\lambda)\log q(x|T(x))$ would be to make the decoder bad at recovering x, which would not encourage $T(x)$ to encode as little information as possible. To prevent this, we use a percentage K of the batches to learn the adversarial decoder $\log q(x|T(x))$, where the encoder is kept frozen. The rest of the batches are used to optimize the full term $-\lambda \log q(y|T(x)) + (1-\lambda)\log q(x|T(x))$, but the gradients for the second term are only propagated to the encoder.

6 Experimental Design

We experiment with the following systems:

- **Proposed.** Our system described in §5. We share the weights between the MT decoder and the adversarial decoder, indicating the language that should be decoded through a special language ID token. Unless otherwise indicated, we use $\lambda = 0.73$ and $K = 0.7$, which performed best in the development set.[7]
- **Round-trip MT.** A baseline that uses two separate MT models to translate into a pivot language and back (see §3).
- **Copy.** A baseline that copies the input text.

We use mBART (Liu et al., 2020) to initialize both our proposed system and round-trip MT, and train them using the same hyper-parameters as in the original work.[8] In both cases, we use the English-French WMT14 dataset (Bojar et al., 2014) as our parallel corpus for training.[9] We report results for two decoding strategies: beam search with a beam size of 5, and top-10 sampling with a temperature of 0.9 (optimized in the development set).[10]

[7]We performed a grid search, where $\lambda \in \{0.7, 0.73, 0.8\}$ and $K \in \{0.7, 0.8\}$, and chose the checkpoint with best iBLEU with $\alpha = 0.7$.

[8]0.3 dropout, 0.2 label smoothing, 2500 warm-up steps, $3e-5$ maximum learning rate, and $100K$ total steps.

[9]We filter the dataset by removing sentence pairs with a source/target length ratio that exceeds 1.5 or are longer than 250 words.

[10]In the case of round-trip MT, we always use beam search to generate the pivot translation, and compare the two approaches to generate paraphrases from it.

Model	Self-BLEU ↓ (diversity)	BLEU ↑ (fidelity)	iBLEU ↑ (combined)
Copy	100.0	23.0	-13.9
MT (beam)	51.1	18.8	-2.17
MT (sampling)	41.4	15.8	-1.36
Ours (beam)	33.0	15.5	0.95
Ours (sampling)	27.3	13.2	**1.05**
Human	18.1	19.8	8.43

Table 1: **Results on the MTC dataset for three baselines (top rows), our two systems, and human performance.** ↓ smaller is better, ↑ larger is better.

We consider two axes when evaluating paraphrases: *fidelity* (the extent to which the meaning of the input text is preserved) and *diversity* (the extent to which the surface form is changed). Following common practice, we use a corpus of gold paraphrases to automatically measure these. More concretely, given the source sentence s, the reference paraphrase r and the candidate paraphrase c, we use BLEU(c, r) as a measure of fidelity, and BLEU(c, s)—known as self-BLEU—as a measure of diversity. An ideal paraphrase system would give us a high BLEU, with as low a self-BLEU as possible. Given that there is generally a tradeoff between the two, we also report iBLEU = α BLEU $-(1-\alpha)$ self-BLEU, which combines both metrics into a single score (Mallinson et al., 2017). Following Hosking and Lapata (2021), we set $\alpha = 0.7$.

For development, we extracted 156 paraphrase pairs from the STS Benchmark dataset (Cer et al., 2017), taking sentence pairs with a similarity score above 4.5. For our final evaluation, we used the Multiple Translations Chinese (MTC) corpus (Huang et al., 2002), which comprises three sources of Chinese journalistic text translated into English by multiple translation agencies. We extract the translations of the first two agencies to obtain an test set of 993 paraphrase pairs, where one is the source and the other the reference paraphrase. The third sentence if kept as an additional paraphrase for estimating human performance.

7 Results

We next report our main results (§7.1), followed by a qualitative analysis (§7.2).

7.1 Main results

We report our main results in Table 1. As it can be seen, our proposed system outperforms all baselines in terms of iBLEU, indicating that it achieves

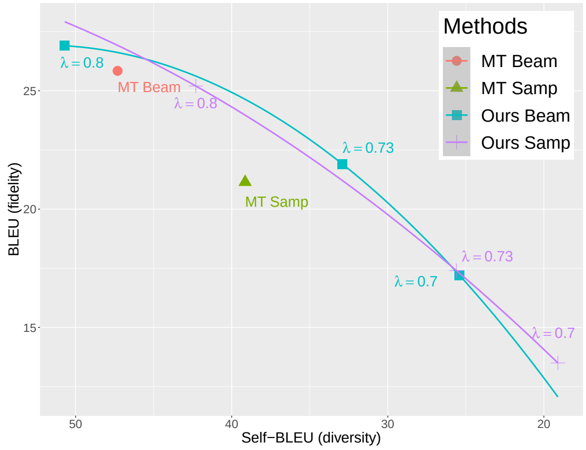

Figure 3: **Effect of varying the λ parameter on the development set.** BLEU in the vertical axis. The horizontal self-BLEU axis is mirrored, so systems toward the top right have the best trade-off between diversity and fidelity.

a better trade-off between diversity and fidelity. This advantage comes from a large improvement in diversity as measured by self-BLEU, at a cost of a small drop in fidelity as measured by BLEU. Both for round-trip MT and our proposed system, beam search does better than sampling in terms of fidelity, at the cost of sacrificing in diversity. Finally, the human reference scores show ample room for improvement in both axes.

While our proposed system achieves the best combined score, our results also show that different approaches behave differently in terms of diversity and fidelity. In practice, it would be desirable to have a knob to control the trade-off between the two, as one may want to favor diversity or fidelity depending on the application. One additional advantage of our approach over round-trip MT is that it offers an adjustable parameter λ to control the trade-off between these two axes. So as to understand the effect of this parameter, we tried different values of it in the development set, and report the resulting curve in Figure 3 together with the MT baselines. BLEU and Self-BLEU scores of the best checkpoints for each λ (0.7,0.73,0.8) and plot the results together with the MT baselines for our systems in Figure 3.

As expected, higher values of λ yield systems that tend to copy more, being more faithful but less diverse. Consistent with our test results, we find that, for a given value of λ, beam search does better than sampling in terms of fidelity, but worse in terms of diversity, yet both decoding strategies can be adjusted to achieve a similar trade-off. More importantly, we observe that both curves are above round-trip MT, the gap being largest for the sampling variant. We can thus conclude that our proposed approach does better than round-trip MT for a comparable trade-off between diversity and fidelity, while offering a knob to adjust this trade-off as desired.

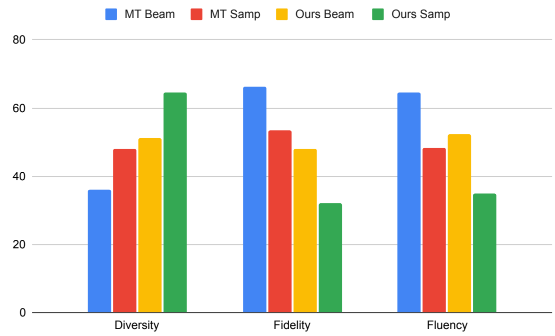

Figure 4: **Human evaluation results (larger is better).** Refer to §7.2 for more details.

7.2 Qualitative analysis

So as to better understand the behavior of our approach in comparison with round-trip MT, we carried out a human evaluation through Amazon Mechanical Turk. Following the setup of Hosking and Lapata (2021), we sample 200 sentences from the MTC corpus and generate a pair of paraphrases for each of them, randomly choosing two systems to generate them. We then ask human evaluators to compare the two sentences according to three criteria: diversity, fidelity and fluency. More details about the judging criteria can be found in Appendix D.

Figure 4 reports the percentage of head-to-head comparisons that each system has won. The results that we obtain are consistent with the trends observed in §7.1. More concretely, we observe that the beam search variant of round-trip MT achieves the best results in terms of fluency and fidelity, but does worst in diversity, indicating a tendency to copy. Our method with beam search does slightly better than the sampling MT variant in terms of diversity and slightly worse in terms of fidelity —indicating a tendency to favor diversity over fidelity— while also being more fluent. Finally, the sampling variant of our method achieves the best diversity, but has the worst fidelity and fluency.

So as to further contrast these results, we manu-

Original	MT (beam)	MT (sampling)	Ours (beam)	Ours (sampling)
The index would fall 3.9% if the sales of vehicles were not included.	The index would fall by 3.9 per cent if vehicle sales were not included.	The Index would fall 3.9% if the vehicle sales were not included, or 3.9% if the vehicle sales were excluded.	The index would decline by 3.9% if the vehicle sales were not included.	The index would decline by 3.9% if vehicles were not included in sales.
Some people worry that this will affect the business of large-sized Canadian enterprises.	There are concerns that this may affect the operations of large Canadian companies.	There is concern that this may affect what large Canadian firms have in their business.	Some may be concerned that this will affect the major Canadian enterprises.	Some people worry that this will impact on the great enterprises in Canada.
These people can set examples and they can have direct influence over the improvement of local human rights conditions and the protection of employees.	These individuals can provide examples and have a direct influence on the improvement of local human rights and employee protection conditions.	These may lead to examples and direct influence on the betterment of local human rights conditions and on the protection of wage earners.	They can provide examples and can directly influence the improvement of local human rights conditions and the protection of employees.	They can provide examples and have direct influence on improving local human rights and the protection of employees' conditions.
The National Youth League said these activities are aimed at showing support and adoration for the state leaders.	The National Youth League stated that these activities are aimed at showing support and admiration to State leaders.	The National Youth League (NDY) stated that these activities are aimed at demonstrating support and admiration for State leaders.	The National League of Youth indicated that these activities are intended to provide support and encourage state leaders.	The National Youth League has stated that such activities are aimed at showing support and admiration for State leaders.

Table 2: Sample paraphrases generated by the different methods.

ally analyzed some paraphrases,[11] and report some examples in Table 2. Just in line with our previous results, we observe that the beam search variant of round-trip MT tends to deviate the least from the original sentence, while the sampling variant of our method generates the most diverse paraphrases (e.g., changing *"sales of vehicles were not included"* to *"vehicles were not included in sales"*). At the same time, we observe that this tendency to improve diversity can cause artifacts like paraphrasing named entities (e.g., changing *"National Youth League"* to *"National League of Youth"*), which can partly explain the drop in fidelity.

8 Conclusions

In this work, we have shown that the implicit similarity function present in round-trip MT is not appropriate in the general case, as it considers sentence pairs that share a single ambiguous translation to be paraphrases. We address this issue by designing an alternative similarity function that requires the entire translation distribution to match, and develop a relaxation of it through the IB method, which we prove to be less susceptible to the problem of confounding translations. We implement this approach through adversarial learning, training an encoder to preserve as much information as possible about the reference translation, while encoding as little as possible about the source. Not only is our approach more principled than round-trip MT, but it is also more efficient at inference, as it does not need to generate an intermediate translation. In addition, it offers a knob to adjust the fidelity-diversity trade-off through the λ parameter, and obtains strong results in our experiments, outperforming round-trip MT.

[11] We randomly sampled 20 sentences from MTC and chose four illustrative examples for Table 2. The 20 random examples are shown in Appendix E.

Acknowledgments

Aitor Ormazabal, Gorka Labaka, Aitor Soroa and Eneko Agirre were supported by the Basque Government (excellence research group IT1343-19 and DeepText project KK-2020/00088) and the Spanish MINECO (project DOMINO PGC2018-102041-B-I00 MCIU/AEI/FEDER, UE). Aitor Ormazabal was supported by a doctoral grant from the Spanish MECD. Computing infrastructure funded by UPV/EHU and Gipuzkoako Foru Aldundia.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 597–604, USA. Association for Computational Linguistics.

Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, ACL '01, page 50–57, USA. Association for Computational Linguistics.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, et al. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the ninth workshop on statistical machine translation*, pages 12–58.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, Hong Kong. Association for Computational Linguistics.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Tom Hosking and Mirella Lapata. 2021. Factorising meaning and form for intent-preserving paraphrasing.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. *CoRR*, abs/1901.03644.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1022–1033, Online. Association for Computational Linguistics.

Shudong Huang, David Graff, and George R. Doddington. 2002. Multiple-translation chinese corpus.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-Guided Controlled Generation of Paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

Nitin Madnani and Bonnie J Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5171–5180. PMLR.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6033–6039, Florence, Italy. Association for Computational Linguistics.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 1999. The information bottleneck method. pages 368–377.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Round-trip MT with restricted sampling

Our formulation in Section 3 considers all possible translations into the pivot language. In practice, some round-trip MT systems use restricted sampling, considering only a subset of Y. For example, ParaNET (Mallinson et al., 2017) takes the K highest probability translations given by beam search. As we show next, the fundamental analysis in Section 3 still holds in that case, and the problem of confounding translations can even be exacerbated by it.

More concretely, using this pivot selection strategy yields the following adjusted paraphrase probability:

$$P(x_p|x_s) = P(x_p) \sum_{y \in \{y_1,\ldots,y_K\}} \frac{P(y|x_s)P(y|x_p)}{ZP(y)}, \tag{12}$$

where $\{y_1, \ldots, y_K\}$ are the top translation candidates and $Z = \sum_{y \in \{y_1,\ldots,y_K\}} P(y|x_s)$. In general, if a subset $S(x_s) \in Y$ of the translation space is considered as pivots, the paraphrase probability will be

$$P(x_p|x_s) = P(x_p) \sum_{y \in S(x_s)} \frac{P(y|x_s)P(y|x_p)}{Z(x_s)P(y)} = \frac{P(x_p)}{Z(x_s)} S'_{MT}(x_p, x_s), \tag{13}$$

where $Z(x_s) = \sum_{y \in S(x_s)} P(y|x_s)$ is a normalizing factor that doesn't depend on the paraphrase x_p, and $S'_{MT}(x_p, x_s)$, is the same similarity function as S_{MT}, with the sum over y restricted to $S(x_s)$.

Using restricted pivot selection strategies such as beam search, top-K sampling, or nucleus sampling (Holtzman et al., 2020) will yield different pivot subsets $S(x_p)$, which will lead to paraphrase probabilities being assigned based on a limited subset of the entire translation distribution. This can exacerbate the issues outlined in Section 3, where the similarity metric can be dominated by a single shared high-probability translation. For example, in the case of translating from a gendered to a genderless language, while the highest probability translation will be genderless, a lower probability candidate might identify the gender, so skipping this translation sampling would increase the similarity of sentences that differ only in gender.

B Characterizing the encoding learned through the IB method

Since we will not learn a perfect minimal sufficient statistic in practice, it is desirable to characterize what the relaxation of S implemented by the IB method can learn.

To that end, we will characterize the kind of encoding T that is allowed by a given γ. Since T is a function of X, we know that $I(X,Y) \geq I(T,Y)$, and thus the condition $I(T,Y) \geq \gamma$ can be rewritten as $I(X,Y) - I(T,Y) \leq \epsilon$, where $\epsilon \geq 0$, which is the form we use throughout this section.

Now, for matters of conditional translation probabilities $P(y|T(x))$, the encoding T can be fully characterized by the equivalence relation it defines on X, where $x_1 \sim x_2$ iff $T(x_1) = T(x_2)$. Two sentences will induce the same conditional translation distribution $P(y|T(x))$ when they are clustered into the same equivalence class by T.

Theorem 3. *Let $\mathcal{S}$ denote the partition of X induced by the encoding T. We denote the elements of a cluster $S \in \mathcal{S}$ by $S = \{x_1^S, ..., x_{m_S}^S\}$. Then, the information loss $I(X,Y) - I(T,Y)$ is given by:*

$$I(X,Y) - I(T,Y) = \sum_{S \in \mathcal{S}} P(x_1^S) KL(P_{Y|x_1^S} || P_{Y|S}) + ... + P(x_{m_S}^S) KL(P_{Y|x_{m_S}^S} || P_{Y|S}), \tag{14}$$

and the translation probabilities conditioned on a cluster are given by the mixture distribution $P(y|T(x)) = P(y|x \in S) = \alpha_1 P(y|x_1^S) + ... + \alpha_{m^S} P(y|x_{m^S}^S)$, where $\alpha_i = \frac{P(x_i^S)}{P(x_1^S) + ... + P(x_{m^S}^S)}$.

The proof can be found in Section C.2. This theorem expresses the information loss of an encoding T, $I(X,Y) - I(T,Y)$, in terms of the KL divergences between the translation distributions of source sentences $P(y|x)$ and the translation distributions given their encodings $P(y|T(x)) = P(y|x \in S)$. Intuitively, if T clusters together two sentences x_1 and x_2 (i.e. $T(x_1) = T(x_2)$ holds), such that $P_{Y|x_1}$ and $P_{Y|x_2}$ are very different, then the mixture distribution $P_{Y|S}$ will be different from both, and thus the information loss will be large.

We will now obtain more intuitive bounds for the information loss. As seen before, the translation distribution given a cluster $P(y|x \in S)$ can be expressed as a mixture of the individual translation distributions for sentences in the cluster:

$$P(y|x \in S) = \frac{P(x_1^S)}{P(x_1^S) + ... + P(x_{m^S}^S)} P(y|x_1^S) + ... + \frac{P(x_{m^S}^S)}{P(x_1^S) + ... + P(x_{m^S}^S)} P(y|x_{m_S}^S). \tag{15}$$

We can also define mixtures of all the distributions $P(y|x_i^S)$ except one, with the same weights as in $P(y|x \in S)$, except for a re-normalization constant. Explicitly, we define:

$$P_j^S(y) = \sum_{i=1, i \neq j}^{m_S} \frac{P(x_i^S) P(y|x_i^S)}{P(x_1^S) + ... + \widehat{P(x_j^S)} + ... + P(x_{m^S}^S)}, \tag{16}$$

where the hat indicates that that element is skipped. Then, we have the following theorem:

Theorem 4. *Let $\mathcal{S}$ be the partition imposed by the encoding function T on X. We denote the elements of a cluster $S \in \mathcal{S}$ by $S = \{x_1^S, ..., x_{m_S}^S\}$. We define the partial mixtures $P_j^S(y)$ as above. Then, if the information loss satisfies $I(X,Y) - I(T,Y) \leq \epsilon$, we have*

$$\sum_{S \in \mathcal{S}} \sum_{i=1}^{m^S} \frac{P(x_i^S)(\beta_i^S)^2}{2} D_1(P_{Y|x_i^S}, P_i^S)^2 \leq \epsilon, \tag{17}$$

where $\beta_j^S = \frac{P(x_1^S) + ... + \widehat{P(x_j^S)} + ... + P(x_{m^S}^S)}{P(x_1^S) + ... + P(x_{m^S}^S)}$ and D_1 is the L^1 norm distance.

The proof can be found in Section C.3. Intuitively, this states that, if the encoding T clusters a set of sentences $x_1, ..., x_n \in S$ together, then the translation distribution for an element $x_i \in S$, $P_{Y|x_i}$, cannot be too far from the mixture of the rest of the distributions $P_{Y|x_j}$, with $j \neq i$.

In the scenario where there are only two sentences in a cluster, $m^S = 2$, we have $P_1 = P_{Y|x_2^S}$ and $P_2 = P_{Y|x_1^S}$, and the inner sum reduces as follows (the derivation is shown in Section C.4):

$$\sum_{i=1}^{m^S} \frac{P(x_i^S)(\beta_i^S)^2}{2} D_1(P_{Y|x_i^S}, P_i^S)^2 = \frac{P(x_1^S)P(x_2^S)}{2(P(x_1^S) + P(x_2^S))} D_1(P_{Y|x_1^S}, P_{Y|x_2^S})^2 \tag{18}$$

Since clustering all the elements of a set S leads to a bigger information loss than only clustering any two elements $x_1, x_2 \in S$, combining Equation 18 and Theorem 4 we obtain Theorem 2 from §4.2 as a corollary:

Theorem 2. *Suppose T is a solution of the IB optimization problem $\min_T I(X,T)$ s.t $I(T,Y) \geq \gamma$, and $\epsilon = I(X,Y) - \gamma$. If $\mathcal{S}$ is the partition on X induced by T, we have:*

$$\sum_{S \in \mathcal{S}} \max_{x_1, x_2 \in S} \frac{P(x_1)P(x_2)}{2(P(x_1)+P(x_2))} D_1(P_{Y|x_1}, P_{Y|x_2})^2 \leq \epsilon, \tag{19}$$

where D_1 is the L_1 norm distance.

This bound is the easiest to interpret intuitively, as it bounds the pairwise distances between the translation distributions of any two sentences that are considered paraphrases by the encoding T. To sum up the results from this section, the information loss allowance when learning with the IB method bounds the L^1 norm distance between the translation distributions of paraphrases. Thus the entire translation distribution is considered when learning paraphrases, potentially alleviating the problems discussed in Section 3

C Proofs

C.1 Proof of Theorem 1

We know that T is a minimal sufficient statistic of Y if and only if the following condition is satisfied:

$$\frac{P(x|y)}{P(x'|y)} \text{ independent of y} \iff T(x) = T(x') \; \forall x, x' \in X \tag{20}$$

Rewriting $P(x|y) = \frac{P(y|x)P(x)}{P(y)}$ and cancelling terms, the LHS becomes:

$$\frac{P_y(x)}{P_y(x')} = \frac{P(y|x)}{P(y|x')}\frac{P(x)}{P(x')}. \tag{21}$$

Since $\frac{P(x)}{P(x')}$ does not depend on y, the entire term will not depend on y if and only if $\frac{P(y|x)}{P(y|x')}$ is independent of y. It is easy to see that the ratio of two distributions of y will be independent of y if and only if they are the exact same distribution, and thus we can conclude that if T is a minimal sufficient statistic of Y then $T(x) = T(x') \iff P(y|x) = P(y|x') \forall y \in Y$, or, equivalently, $T(x) = T(x') \iff S(x,x') = 1$.

Thus, we have

$$\begin{aligned} P(x_p|T(x_s) &= P(X = x_p|T(X) = T(x_s)) \\ &= P(X = x_p|S(X, x_s) = 1) \\ &= \frac{P(X = x_p, S(X, x_s) = 1)}{P(S(X, x_s) = 1)} \\ &= \frac{P(X = x_p)P(S(X, x_s) = 1)|X = x_p)}{P(S(X, x_s) = 1)} \\ &= \frac{P(x_p)S(x_p, x_s)}{Z}, \end{aligned} \tag{22}$$

where $Z = P(S(X, x_s) = 1)$ is the normalizer that does not depend on x_p, as we wanted to prove. □

C.2 Proof of Theorem 3

We first expand the information loss:

$$\begin{aligned} &I(X,Y) - I(T,Y) \\ &\quad = \mathbb{E}_X KL(P_{Y|X}||P_Y) \\ &\quad - \mathbb{E}_T KL(P_{Y|T}||P_Y) = \sum_x P(x) \\ &\quad \Big[\sum_y P(y|x)log(P(y|x)) \\ &\quad - P(y|x)log(P(y))\Big] \\ &\quad - \sum_x P(x)\Big[\sum_y P(y|T(x))log(P(y|T(x))) \\ &\quad - P(y|T(x))log(P(y))\Big] \end{aligned} \tag{23}$$

Now, for matters of conditional translation probabilities $P(y|T(x))$, the encoding T can be fully characterized by the equivalence class it defines on X, where $x_1 \sim x_2$ iff $T(x_1) = T(x_2)$. We let $\mathcal{S}$ denote the partition on X induced by this equivalence class. Then, we can rewrite:

$$
\begin{aligned}
&I(X,Y) - I(T,Y) \\
&= \sum_x P(x)\Big[\sum_y P(y|x)log(P(y|x)) \\
&- \ P(y|x)log(P(y))\Big] \\
&- \sum_x P(x)\Big[\sum_y P(y|T(x))log(P(y|T(x))) \\
&- P(y|T(x))log(P(y))\Big] \\
&= \sum_{S\in\mathcal{S}}\sum_{x\in S} P(x)\Big[\sum_y P(y|x)log(P(y|x)) \\
&- P(y|x)log(P(y))\Big] \\
&- \sum_{S\in\mathcal{S}}\sum_{x\in S} P(x)\Big[\sum_y P(y|x\in S)log(P(y|x\in S)) \\
&- P(y|x\in S)log(P(y))\Big]
\end{aligned}
\tag{24}
$$

For a certain $S \in \mathcal{S}$, we denote its elements by $S = \{x_1^S, ..., x_{m_S}^S\}$. Then, we have

$$
\begin{aligned}
P(y|x\in S) &= \frac{P(y, x\in S)}{P(x\in S)} \\
&= \frac{P(y, x_1^S) + ... + P(y, x_{m_S}^S)}{P(x_1^S) + ... + P(x_{m_S}^S)} \\
&= \alpha_1^S P(y|x_1^S) + ... + \alpha_{m_S}^S P(y|x_{m_S}^S),
\end{aligned}
\tag{25}
$$

where $\alpha_i^S = \frac{P(x_i^S)}{P(x_1^S)+...+P(x_{m_S}^S)}$. We also define $\beta^S = P(x_1^S) + ... + P(x_{m_S}^S)$. Now, we can rewrite the first expression of the RHS in Equation 24:

$$
\begin{aligned}
&\sum_{S\in\mathcal{S}}\sum_{x\in S} P(x)\sum_y \big[P(y|x)log(P(y|x)) \\
&- P(y|x)log(P(y))\big] \\
&= \sum_{S\in\mathcal{S}}\Big[P(x_1^S)\sum_y P(y|x_1^S)log(P(y|x_1^S)) + ... \\
&+ P(x_{m_S}^S)\sum_y P(y|x_{m_S}^S)log(P(y|x_{m_S}^S))\Big] \\
&- \Big[\sum_y P(x_1^S)P(y|x_1^S)log(P(y)) + ... \\
&+ P(x_{m_S}^S)P(y|x_{m_S}^S)log(P(y))\Big] \\
&= \sum_{S\in\mathcal{S}}\Big[P(x_1^S)\sum_y P(y|x_1^S)log(P(y|x_1^S)) + ... \\
&+ P(x_{m_S}^S)\sum_y P(y|x_{m_S}^S)log(P(y|x_{m_S}^S))\Big] \\
&- \Big[\sum_y \beta^S\alpha_1^S P(y|x_1^S)log(P(y)) + ... \\
&+ \beta^S\alpha_{m_S}^S P(y|x_{m_S}^S)log(P(y))\Big] \\
&= \sum_{S\in\mathcal{S}}\Big[P(x_1^S)\sum_y P(y|x_1^S)log(P(y|x_1^S)) + ... \\
&+ P(x_{m_S}^S)\sum_y P(y|x_{m_S}^S)log(P(y|x_{m_S}^S))\Big] \\
&- \Big[\beta^S\sum_y P(y|x\in S)log(P(y))\Big]
\end{aligned}
\tag{26}
$$

And now we rewrite the second term in the RHS of Equation 24:

$$
\begin{aligned}
&\sum_{S\in\mathcal{S}}\sum_{x\in S}P(x)\sum_{y}\left[P(y|x\in S)log(P(y|x\in S))\right.\\
&\left.-P(y|x\in S)log(P(y))\right]\\
&=\sum_{S\in\mathcal{S}}\left[P(x_1^S)\sum_{y}P(y|x\in S)log(P(y|x\in S))\right.\\
&\left.+P(x_{m_S}^S)\sum_{y}P(y|x\in S)log(P(y|x\in S))\right]\\
&-\left[P(x_1^S)\sum_{y}P(y|x\in S)log(P(y))\right.\\
&\left.+P(x_{m_S}^S)\sum_{y}P(y|x\in S)log(P(y))\right]\\
&=\sum_{S\in\mathcal{S}}\left[\beta^S\sum_{y}P(y|x\in S)log(P(y|x\in S))\right]\\
&-\left[\beta^S\sum_{y}P(y|x\in S)log(P(y))\right]\\
&=\sum_{S\in\mathcal{S}}\left[P(x_1^S)\sum_{y}P(y|x_1^S)log(P(y|x\in S))\right.\\
&\left.+P(x_{m_S}^S)\sum_{y}P(y|x_{m_S}^S)log(P(y|x\in S))\right]\\
&-\left[\beta^S\sum_{y}P(y|x\in S)log(P(y))\right],
\end{aligned}
\tag{27}
$$

where we have used Equation 25 in the last equality.

Substituting (26) and (27) into (24), we get:

$$
\begin{aligned}
&I(X,Y)-I(T,Y)\\
&=\sum_{S\in\mathcal{S}}\left[P(x_1^S)\sum_{y}P(y|x_1^S)log(P(y|x_1^S))+...\right.\\
&\left.+P(x_{m_S}^S)\sum_{y}P(y|x_{m_S}^S)log(P(y|x_{m_S}^S))\right]\\
&-\sum_{S\in\mathcal{S}}\left[P(x_1^S)\sum_{y}P(y|x_1^S)log(P(y|x\in S))\right.\\
&\left.+P(x_{m_S}^S)\sum_{y}P(y|x_{m_S}^S)log(P(y|x\in S))\right]\\
&=\sum_{S\in\mathcal{S}}P(x_1^S)\sum_{y}\left[P(y|x_1^S)log(P(y|x_1^S))\right.\\
&\left.-P(y|x_1^S)log(P(y|x\in S))\right]+...\\
&+P(x_{m_S}^S)\sum_{y}\left[P(y|x_{m_S}^S)log(P(y|x_{m_S}))\right.\\
&\left.-P(y|x_{m_S}^S)log(P(y|x\in S))\right]\\
&=\sum_{S\in\mathcal{S}}P(x_1^S)KL(P_{Y|x_1^S}||P_{Y|S})+...\\
&+P(x_{m_S}^S)KL(P_{Y|x_{m_S}^S}||P_{Y|S})
\end{aligned}
\tag{28}
$$

As we wanted to show. □

C.3 Proof of Theorem 4

By Theorem 3, it is enough to show that

$$
\begin{aligned}
&\sum_{S\in\mathcal{S}}P(x_1^S)KL(P_{Y|x_1^S}||P_{Y|S})\\
&\qquad+...+P(x_{m_S}^S)KL(P_{Y|x_{m_S}^S}||P_{Y|S})\\
&\qquad\geq\sum_{S\in\mathcal{S}}\sum_{i=1}^{m^S}\frac{P(x_i^S)(\beta_i^S)^2}{2}D_1(P_{Y|x_i^S},P_i^S)^2
\end{aligned}
\tag{29}
$$

For that, it is enough to show that

$$
\begin{aligned}
&P(x_i^S)KL(P_{Y|x_i^S}||P_{Y|S})\\
&\qquad\geq\frac{P(x_i^S)(\beta_i^S)^2}{2}D_1(P_{Y|x_i^S},P_i^S)^2
\end{aligned}
\tag{30}
$$

for every i. Now, by Pinsker's inequality, for a given i, we have that

$$
\begin{aligned}
&P(x_i^S)KL(P_{Y|x_i^S}||P_{Y|S}) = P(x_i^S) \\
&\quad KL(P_{Y|x_i^S}||\alpha_1 P_{Y|x_1^S} + ... + \alpha_{m^S} P_{Y|x_{m^S}^S}) \\
&\quad \geq \frac{P(x_i^S)}{2} D_1(\alpha_1 P_{Y|x_1^S} + ... \\
&\quad + \alpha_{m^S} P_{Y|x_{m^S}^S}, P_{Y|x_i^S})^2 \\
&\quad = \frac{P(x_i^S)}{2(P(x_1^S) + ... + P(x_{m^S}^S))^2} ||_1(P(x_1)P_{Y|x_1^S} + ... \\
&\quad + P(x_{m^S})P_{Y|x_{m^S}^S} \\
&\quad - (P(x_1^S) + ... + P(x_{m^S}^S))P_{Y|x_i^S}||^2 \\
&\quad = \frac{P(x_i^S)}{2(P(x_1^S) + ... + P(x_{m^S}^S))^2} \\
&\quad ||_1(P(x_1^S)P_{Y|x_1^S} + ...\widehat{P(x_i^S)P_{Y|x_i^S}} + ... \\
&\quad + P(x_{m^S}^S)P_{Y|x_{m^S}^S} \\
&\quad - (P(x_1^S) + ... + \widehat{P(x_i^S)} + ... + P(x_{m^S}))P_{Y|x_i^S}||^2 \\
&\quad = \frac{P(x_i^S)(P(x_1^S) + ... + \widehat{P(x_i^S)} + ... + P(x_{m^S}))^2}{2(P(x_1^S) + ... + P(x_{m^S}^S))^2} \\
&\quad ||_1 P_i^S - P_{Y|x_i^S}||^2 \\
&\quad = \frac{P(x_i^S)(\beta_i^S)^2}{2} D_1(P_i^S(Y), P_{Y|x_i^S})^2
\end{aligned} \tag{31}
$$

where the hat represents that element of the sum being skipped, as we wanted to show. □

C.4 Derivation of Equation 18

$$
\begin{aligned}
&\sum_{i=1}^{m^S} \frac{P(x_i^S)(\beta_i^S)^2}{2} D_1(P_{Y|x_i^S}, P_i^S)^2 \\
&= \left[\frac{P(x_1^S)(\beta_1^S)^2}{2} + \frac{P(x_2^S)(\beta_2^S)^2}{2}\right] D_1(P_{Y|x_1^S}, P_{Y|x_2^S})^2 \\
&= \left[\frac{P(x_1^S)P(x_2^S)^2}{2(P(x_1^S) + P(x_2^S))^2} + \frac{P(x_2^S)P(x_1^S)^2}{2(P(x_1^S) + P(x_2^S))^2}\right] \\
&D_1(P_{Y|x_1^S}, P_{Y|x_2^S})^2 \\
&= \left[\frac{P(x_1^S)P(x_2^S)(P(x_1^S) + P(x_2^S))}{2(P(x_1^S) + P(x_2^S))^2}\right] \\
&D_1(P_{Y|x_1^S}, P_{Y|x_2^S})^2 \\
&= \frac{P(x_1^S)P(x_2^S)}{2(P(x_1^S) + P(x_2^S))} D_1(P_{Y|x_1^S}, P_{Y|x_2^S})^2
\end{aligned} \tag{32}
$$

D Human evaluation judging criteria

We ask human evaluators to compare systems on three different dimensions, according to the following instructions:

Meaning. Which of the paraphrases better preserves the meaning of the original, without adding or losing information?

Surface similarity. Which of the paraphrases is more similar compared to the original, using more similar phrasing or words? You should chose the text using more similar words or phrasing, regardless of meaning.

Fluency. Which text is a more fluent English sentence? You should choose the sentence that contains the least grammatical mistakes, and sounds more natural.

E Full paraphrase sample

We present the full list of 20 sampled paraphrases in Table 3.

Original	MT (beam)	MT (sampling)	Ours (beam)	Ours (sampling)
As of August 30th, the city had allocated a labor force of 158,300 per day in the project, with the aggregate labor contribution amounting to 936,500, and had completed 730,000 cubic meters of earth and stone.	As at 30 August, the city had allocated 158,300 persons per day to the project, with an overall contribution of 936,500 persons, and had completed the construction of 730,000 cubic metres of land and stone.	It had assigned 158,300 employees per day to the project as of 30 August, with a total staff contribution of 936,500, and it had concluded the construction of 730,000 cubic metres of earth and rock.	As of 30 August, the city had employed 158,300 persons per day for the project, a total of 936,500 of whom had been allocated for the construction of 230,000 square meters of earth and stone.	On 30 August, the City had a staff of 158,300 people per day in the project, a total of 936,500 of whom had been provided for the construction of 330,000 square foot of earth andstone work.
The United States and North Korea are scheduled to hold talks on Friday regarding US access to a suspected underground nuclear site in the vicinity of Pyongyang.	The United States and North Korea are expected to hold talks on Friday on United States access to a suspected underground nuclear site near Pyongyang.	The United States and North Korea are scheduled for talks Friday on the US access to a suspected underground nuclear site near Pyongyang.	The United States and North Korea are scheduled to hold talks Friday on the access of the United States to a suspected nearby Pyongyang ground nuclear site.	The United States and North Korea are scheduled to hold talks on Friday on United States access to a suspected nearby Pyongyang nuclear site.
For long, Xi'ning has made insufficient investment in the construction of urban infrastructure facilities, with the total investment made being only about RMB400 million during the 46-year period between the founding of the New China to 1995. As a result, the backward and underdeveloped infrastructure facilities have restricted the city's economic development.	Xi'ning has for a long time not invested sufficiently in the construction of urban infrastructure, with a total investment of only about RMB 400 million over the 46-year period between the founding of New China and 1995, as a result of which lagging and underdeveloped infrastructure has limited the city's economic development.	Xi'ning had for a long time not invested enough in construction of urban infrastructure, with total investment amounting only to around RMB 400m during the 46-year period between the founding of New China and 1995. Therefore, backward and underdeveloped infrastructure had impeded economic development of the city.	Since long Xi'ning's investment in urban infrastructure has been insufficient, with a total investment of only 400 million cubic metres in total, for 46 years between the founding of the New China in 1995, thus limiting the development of infrastructure in the underdeveloped and underdeveloped areas.	Since long Xi'ning has not made adequate investment in urban infrastructure, with an overall investment of only \$400 million MB over 46 years from the creation of the new China to 1995, thus restricting the development of urban infrastructure, both underdeveloped and underdeveloped.
Japan, Australia, New Zealand and South Korea Expresses Support, saying that the U.S. Has No Other Choice	Japan, Australia, New Zealand and South Korea expressed their support, stating that the United States had no other choice.	Japan, Australia, New Zealand and South Korea show their support, affirming that the United States does not have no Alternative	Japan, Australia, New Zealand and South Korea expressed their support, stating that the United States had no other choice.	Japan, Australia, New Zealand and South Korea expressed support, stating that the United States had no other choice.
Chernomyrdin also pointed out that there were also many problems in Russia last year, such as the poor performance of taxation, investment plans yet to be completed, reduced surplus in foreign trade, and ineffective fiscal and financial measures taken by the government.	Chernomyrdin also pointed out that there were also many problems in Russia last year, such as poor fiscal performance, investment plans that had not yet been completed, the reduction of the foreign trade surplus and ineffective fiscal and financial measures taken by the government.	Chernomyrdin also stressed that, in Russia, last year, there were many problems too. These included a deteriorating fiscal performance, investment plans that have yet to be completed, the reduction of foreign trade surpluses, and inefficient fiscal and financial measures taken by the Government, etc.	Chernomyrdin also noted that there were many problems in Russia last year, such as poor taxation performance, plans for investment still to be completed, the reduction in foreign trade surplus, and inadequate fiscal and financial measures taken by the Government.	Chernomyrdin further noted that there were many further problems in Russia last year, such as weak taxation performance, plans for further investment still to be completed, the reduction in foreign trade surplus, and ineffective fiscal and financial measures taken by the Government.
Based on the plan, the GDP in Russia next year is to increase by 2%, and the inflation rate is to go down to 5% to 8%.	According to the plan, Russia's GDP will rise by 2 per cent next year and the inflation rate will rise from 5 per cent to 8 per cent.	According to the plan, Russian GDP will rise 2 per cent next year and the rate of inflation will be reduced from 5 per cent to 8 per cent.	According to this plan, GDP in Russia will increase by 2% next year and the inflation rate will drop from 5% to 8%.	Under the plan, the GDP in Russia will increase 2 per cent next year and the un-inflation rate increase from 5% to 8%.

Zuo Zhongmo, deputy secretary-general of the Conference, said, "This is not just an issue of agriculture. These reclaimed lands can serve the general development purposes of various sectors, including forestry, industry and tourism."	Zuo Zhongmo, Under-Secretary-General of the Conference, said: “This is not just an agriculture issue, these recovered lands can serve the overall development goals of various sectors, including forestry, industry and tourism.	Zuo Zhongmo, Deputy Secretary-General of the Conference, said, "It is not just an agriculture issue; the land recovered from them can serve the overall development goals of different sectors including forestry, industry and tourism."	It is Zuo Zhongmo, Under-Secretary-General of the Conference, who said: It is not just about agriculture; the reclaimed lands can be used for general development in various sectors, including forestry, industry and tourism."	The Assistant Secretary-General of the Conference, Zuo Zhongmo, said, It is not just about agriculture, as the reclaimed lands can be used for general development purposes from various sectors, including forestry, industry and tourism.
In the United States, California and other southern states were flooded at the beginning of this year, followed by a drought in many places in the south.	In the United States, California and other southern states were flooded early this year, followed by droughts in many parts of the South.	In the United States of America, CaliFORNA and other southern states were flooded early this year, followed by drought in many parts of the south.	In the United States, California and other southern states, floods occurred in early this year, followed by drought in many areas of the south.	Inondations in California, and other southern States in early this year, followed by droughts in many endroits in the south.
Meanwhile, the US Congress is discussing whether or not to approve the Protocol reached in the Kyoto Conference in Japan.	In the meantime, the United States Congress is discussing whether or not to approve the Protocol concluded at the Kyoto Conference in Japan.	Meanwhile, US Congress debates whether or not to accede to the Protocol agreed at the Kyoto Conference in Japan.	At the same time, the United States Congress is examining whether or not to approve the Protocol at the Kyoto Conference in Japan.	At the same time, the United States Congress is considering whether or not to approve the Protocol made at the Kyoto Conference in Japan.
The index would fall 3.9% if the sales of vehicles were not included.	The index would fall by 3.9 per cent if vehicle sales were not included.	The Index would fall 3.9% if the vehicle sales were not included, or 3.9% if the vehicle sales were excluded.	The index would decline by 3.9% if the vehicle sales were not included.	The index would decline by 3.9% if vehicles were not included in sales.
According to the company, in the coming five years, the company will make an additional investment of US$90 million, with an anticipated annual output value of US$300 million.	According to the company, over the next five years, it will make an additional investment of US$90 million, with a projected annual production value of US$300 million.	According to the company, it will make an additional $90 million US over the next five years with a planned annual productivity value of $300 million US.	According to the company, over the next five years it will invest an additional $90 million in the U.S., its expected annual output of $300 million.	According to the company, it will invest $90 million in the next five years, with its expected annual output of approximately $300 million.
Some people worry that this will affect the business of large-sized Canadian enterprises.	There are concerns that this may affect the operations of large Canadian companies.	There is concern that this may affect what large Canadian firms have in their business.	Some may be concerned that this will affect the major Canadian enterprises.	Some people worry that this will impact on the great enterprises in Canada.
A Majority of Hong Kong Residents Decline to Consider Themselves as Chinese	A majority of Hong Kong residents decline to regard themselves as Chinese	The Hong Kong City of Hong Kong’s minority people are not recognising themselves as Chinese	The vast majority of Hong Kong residents feel they are being referred as Chinese	The vast majority of Hong Kong residents have become disenfranchised as Chinese
He said the 83-year-old woman has been hospitalized for overshock.	He said that the 83-year-old woman had been hospitalized for overheating.	He said that the 83-year-old woman had been hospitalized for overheat.	He stated that a 83-year-old woman had been hospitalized for a headache.	He indicated that a 83-year-old woman had been hospitalized for a head injuries.
However, statistics released by the Immigration Bureau showed that although there were 11,978 new British immigrants coming to Australia between 1996 and 1997, 3,737 people left the country during the same period.	However, statistics published by the Immigration Bureau show that, although there were 11,978 new British immigrants to Australia between 1996 and 1997, 3,737 people left the country during the same period.	However, Immigration Bureau statistics show that although there were 11,978 new British immigrants to Australia from 1996 to 1997, 3,737 individuals left the country during that period.	However, the statistics of the Office of Immigration show that, although 11,978 new British immigrants arrived in Australia between 1996 and 1997, 3,737 had left the country during the same period.	However, the figures from the Immigration Bureau show that, although 11,978 new British immigrants had entered Australia between 1996 and 1997, 3,737 had left the country during the same period.

(Reuters report from Tokyo)Japanese Finance Minister Kiichi Miyazawa was pressured not to quit office yesterday.	Japan’s Finance Minister, Kiichi Miyazawa, was forced yesterday not to leave his post.	The finance minister in Japan, Kiichi Miyazawa, was forced to leave the post yesterday.	(Tokyo report) Japan’s Minister of Finance Kiichi Mi Theawa was pressured not to quit yesterday.	(Tokyo report) The Japanese Finance Minister Kiichi Mi ichiawa was pressured on not to leave yesterday.
US Admitted Hundreds of Deaths of Iraqi Civilians in Air Strike and UK Reporter Claimed the Target Being Definitely Non-military	US Admitted Hundreds of Deaths of Iraqi Civilians in Air Strike and UK Reporter Claimed the Target Being Definitely Non-military	U.S. Admitted Hundreds of Deaths of Iraqi Civilians in Air Strike and UK Reporter Claimed the Target Being Definitely Non- Military	The US killed hundreds of Iraqi civilians in an air strike, and a reporter in the UK said the target was essentially non-military.	Several hundred dead of Iraqi civilians in air strikes; a UK reporter said that its specific goal is, essentially, non-military.
During the Eighth Five-Year Plan Period (from 1991 to 1995), township enterprises in Fujian Province contributed an aggregate total of RMB18.56 billion in tax, and achieved a total of RMB105.5 billion worth of export commodities.	During the eighth period of the five-year plan (1991-1995), municipal enterprises in Fujian Province paid taxes totalling RMB 18.56 billion and exported products totalling RMB 105.5 billion.	During the eighth period of the Five-Year Plan (from 1991 to 1995), Fujian Provincial Municipal Enterprises had paid taxes totalling RMB 18.56 billion and exported revenues totalling RMB 105.5 billion.	During the eighth five-year plan (from 1991 to 1995), businesses in the townships of Fujian Province contributed a total of \$18.56 billion in tax contributions, \$105.5 billion in commodities for export.	During the eighth five-year plan (from 1991 to 1995), businesses in the townships of Fujian Province contributed tax contributions totalling \$118.56 billion in 1991-95, as well as \$1005.5 billion in commodities for export.
In May this year, Dole admitted using Viagra on a trial basis, and gave high remarks to the drug after use, describing it as "a magic drug."	In May of this year, Dole admitted to using Viagra on an experimental basis and commented very positively on the drug after its use, describing it as "a magical drug."	In May this year, Dole admitted that he uses Viagra as an experimental patient and provided very positive feedback on the drug after its use describing it "a magical medicine".	In May of this year, Dole accepted the use of marijuana for trials and reported strong post-treatment remarks, describing it as a magical drug.	In May of this year, Dole recognized the use of pesticide in trials and had reported very good after-treatment, describing it as the “ magical drug”.
These people can set examples and they can have direct influence over the improvement of local human rights conditions and the protection of employees.	These individuals can provide examples and have a direct influence on the improvement of local human rights and employee protection conditions.	These may lead to examples and direct influence on the betterment of local human rights conditions and on the protection of wage earners.	They can provide examples and can directly influence the improvement of local human rights conditions and the protection of employees.	They can provide examples and have direct influence on improving local human rights and the protection of employees’ conditions.

Table 3: Sample paraphrases generated by the different methods.