What Can Unsupervised Machine Translation Contribute to High-Resource Language Pairs?

Kelly Marchisio
Johns Hopkins University
kmarc@jhu.edu

Markus Freitag  David Grangier
Google Research
{freitag, grangier}@google.com

Abstract

Whereas existing literature on unsupervised machine translation (MT) focuses on exploiting unsupervised techniques for low-resource language pairs where bilingual training data is scarce or unavailable, we investigate whether unsupervised MT can also improve translation quality of high-resource language pairs where sufficient bitext does exist. We compare the style of correct translations generated by either supervised or unsupervised MT and find that the unsupervised output is less monotonic and more natural than supervised output. We demonstrate a way to combine the benefits of unsupervised and supervised MT into a single system, resulting in better human evaluation of quality and fluency. Our results open the door to discussions about the potential contributions of unsupervised MT in high-resource settings, and how supervised and unsupervised systems might be mutually-beneficial.

1 Introduction

Unsupervised machine translation (MT) uses monolingual data alone to learn to translate. Popular techniques include “mapping”-based methods which induce and iteratively refine a bilingual dictionary from monolingual data (e.g. Artetxe et al., 2019; Lample et al., 2018b; Conneau et al., 2018), and methods which finetune a model pretrained with a language-modeling objective (e.g. Conneau and Lample, 2019; Song et al., 2019). Unlike supervised MT, which is trained on parallel data, unsupervised MT can be trained with natural data alone.

Although traditionally motivated by language pairs which lack bilingual training data, it is worth asking whether unsupervised MT could benefit other language pairs as well. Recent unsupervised MT systems can reach reasonable translation quality under clean and controlled data conditions (Song et al., 2019; Artetxe et al., 2019; Conneau and Lample, 2019), and could bring alternative translations to language pairs with ample clean parallel data. Whereas existing unsupervised MT work strives to provide translations for low-resource pairs for which little or no bilingual training data exists, our primary goal is to investigate the potential contribution of unsupervised MT to high-resource language pairs. Our exploration of this research area focuses on English→German for which abundant bilingual training examples exist.

Our main contributions are:
• We observe a systematic difference in style between the output of supervised and unsupervised MT systems of similar quality.
• We demonstrate how to combine unsupervised MT with supervised MT to improve quality, naturalness, and fluency of translation based on human evaluation.
• We develop a new metric for measuring monotonicity in translation.
• Our results provoke interesting questions about potential contributions of unsupervised MT in high-resource scenarios, and indicate that combining supervised and unsupervised setups may contribute to a system better than either creates alone.

The paper is structured as follows: We discuss related work in §2. In §3, we introduce the dataset, model details and evaluation setups which will be used in our study. In §4, we characterize the differences between the output of unsupervised and supervised NMT. For fair comparison, we build two systems with similar translation quality. Based on the findings, we train a single system which benefits from the complementarity of supervised and unsupervised methods in §6. We present ablation experiments in §5, and summarize in §7.

* Work completed at Google Translate Research.
2 Related Work

Unsupervised MT Two major paradigms for unsupervised MT are finding a linear transformation to align two monolingual embedding spaces (Lample et al., 2018a,b; Conneau et al., 2018; Artetxe et al., 2018, 2019), and pretraining a bilingual or multilingual language model before finetuning on the translation task (Conneau and Lample, 2019; Song et al., 2019; Liu et al., 2020). In this work, we study the Masked Sequence-to-Sequence Pretraining (MASS) unsupervised language model pretraining paradigm of Song et al. (2019).

Using Monolingual Data in MT Back-translation (BT) (Sennrich et al., 2016) is widely-used to exploit knowledge in monolingual data. Quality gains can be attained by adding a small number of parallel sentences to an unsupervised system (semi-supervised, e.g. Artetxe et al. (2018)). Siddhant et al. (2020) combine multilingual supervised training with the MASS objective on monolingual data to boost performance for various languages and in zero-shot translation.

Source Artifacts in Translated Text Because supervised MT is trained ideally on human-generated translation, characteristics of human translation can affect the characteristics of machine-generated translation from supervised systems. Sometimes dubbed “translationese”, human translation includes source language artifacts (Koppel and Ordan, 2011) as well as source-independent artifacts—Translation Universals (Mauranen and Kujamäki, 2004). The professional translation community studies systematic biases inherent to translated texts (Baker, 1993; Selinker, 1972), and biases resulting from interference from source text (Toury, 1995). In MT, Freitag et al. (2019, 2020) point at these patterns as a source of mismatch between BLEU (Papineni et al., 2002) and human evaluation measures of quality, raising concerns that overlap-based metrics may reward hypotheses with the characteristics of translated language more than those with natural language. Vanmassenhove et al. (2019, 2021) note loss of linguistic diversity and richness from MT, and Toral (2019) note related effects even after human post-editing. The impact of translated text on human evaluation has also been studied (Toral et al., 2019; Zhang and Toral, 2019; Graham et al., 2019; Fomicheva and Specia, 2016; Ma et al., 2017), as has the impact in training data (Kurokawa et al., 2009; Lemersky et al., 2012; Bogoychev and Sennrich, 2019; Riley et al., 2020).

3 Experimental Setup

3.1 Data

Training Sets We use the English→German training data from WMT2018 (Bojar et al., 2018) in our main experiments (§6). We use News Crawl 2007-17 for monolingual data, which contains 165 million and 226 million sentences in English and German, respectively after deduplication. As bilingual training data, we use News Commentary v13, Europarl v7, Common Crawl and EU Press Release. We deduplicate and filter out pairs with > 250 tokens in either language or length ratio over 1.5. This results in 5.2 million parallel sentence pairs.

Development and Test Sets Newstest2017 and newstest2018 are used for development and newstest2019 for test, abbreviated nt17, nt18, and nt19, respectively. We also use paraphrased test references from Freitag et al. (2020), hereafter called “nt18p” and “nt19p”. These additional references are provided for data with original English source and measure system quality without favoring sentences that use the same structure as the source.

Measuring Word Reordering Research into word reordering models is well-studied because word reordering once formed a critical part of MT systems in the days of statistical MT. See Bisazza and Federico (2016) for a review. Other work has looked into metrics for measuring word reordering in translation (e.g. Birch et al., 2008, 2009, 2010). Wellington et al. (2006) and Fox (2002) use POS tags within the context of parse trees, and Fox (2002) measure the similarity of French and English with respect to phrasal cohesion by calculating alignment crossings using parse trees. Most similar to our work, Birch (2011) view simplified word alignments as permutations and compare distance metrics over these to quantify the amount of reordering that has taken place. Birch and Osborne (2011) present LRScore which interpolates a reordering metric with a lexical translation metric.

1 https://github.com/google/wmt19-paraphrased-references
sentence. BLEU on the paraphrase test set are generally much lower than on original test sets, and small score differences (such as 0.3) can indicate real quality differences. As suggested by Freitag et al. (2019), we report BLEU scores split by original language. This results into src-orig and tgt-orig halves for each test set. We report SacreBLEU (Post, 2018) throughout the paper.

3.2 Part-of-Speech Tagging
We use part-of-speech taggers in some of our experiments. We use universal dependencies implemented in spaCy, and spaCy’s language-specific fine-grained POS tags for German from the TIGER Corpus (Albert et al., 2003; Brants et al., 2004).

3.3 Models
The unsupervised MT model used in this work is a MASS transformer with the hyperparameters of Song et al. (2019), trained on the News Crawl corpora. It is hereafter called “Unsup”.

We train supervised systems using the transformer-big (Vaswani et al., 2017) architecture as implemented in Lingvo (Shen et al., 2019). We use 32k subword units, train models for roughly 500k updates, and choose checkpoints based on validation performance on nt18.

To investigate the differences between the translation approaches, we train 2 language models (LMs) on different types of training data and calculate the probabilities on the translations generated by either the supervised or the unsupervised approach. First, we train a LM on the monolingual German News Crawl dataset with a decoder-only transformer, hereafter called the “natural text LM” (nLM). We train another LM on machine translated sentences which we call the “translated text LM” (tLM). We generate the training corpus by translating the monolingual English News Crawl dataset into German with a German→English transformer-big model trained on the WMT18 bitext.

3.4 Human Evaluations
First, we evaluate quality using direct assessment (as done in WMT). Second, we run two side-by-side evaluations to measure fluency and quality preference between two systems. Each evaluation includes 1,000 items. We hired a pool of 12 professional translators as they are more reliable than crowd workers (Toral, 2020; Freitag et al., 2021). We evaluate the originally-English sentences corresponding to the official WMT-19 en→de test set, because Zhang and Toral (2019) show that a source that is in reality translated text should not be used for human evaluation.

Direct Assessment Quality Quality is evaluated with the template from the WMT 2019 evaluation campaign. Human translators assess a given translation by how adequately it expresses the meaning of the corresponding source sentence on a 0-100 scale. Unlike WMT, we report the average rating and do not normalize the scores.

Side-by-side Quality Raters see a source sentence with two alternative translations, and rate each on a 6-point scale (See Tables 3, 9).

Side-by-side Fluency Raters assess two alternative German sentences without the source, and rate each on a 6-point scale (See Table 10).

4 Differences between Supervised and Unsupervised Outputs

4.1 Model Performance
We investigate the differences between the output of unsupervised and supervised NMT models. To make the comparison fair, we consider systems of similar quality so that quality of the output does not confound analyses and we still get an impression of the differences between the translation approaches.

|          | orig-de | orig-en | BLEU |
|----------|---------|---------|------|
| SupWMT14 | 40.9    | 34.9    | 44.6 | 12.1 |
| SupNC    | 29.2    | 21.1    | 34.0 | 9.3  |
| Unsup    | 30.1    | 27.1    | 30.9 | 9.6  |

Table 1: SacreBLEU on nt18 and nt18p (BLEUp). Nt18 score is also reported for sentences originally written in German (orig-de) or English (orig-en).

As unsupervised systems are expected to under-perform supervised systems given the same training data, we use an amended training corpus to achieve systems of similar quality. We therefore consider a supervised setup with less training data. We train a supervised system on News Commentary ("SupNC") only and a supervised system trained on the full WMT14 training set ("SupWMT14"). Even...
though the supervised system is below state-of-the-art, these experiments can help to understand if and how unsupervised output is better/different when the accuracy is similar to the supervised model.

Table 1 summarizes BLEU scores. SupNC and Unsup show notable differences between orig-de and orig-en sides of the test set despite having similar BLEU overall. Recalling that we translate from English to German, the fact that Unsup has higher BLEU than SupNC when the target-side is natural text (orig-de) may suggest that its output is more natural-sounding because it better matches text originally written in German. This first insight makes sense, as the unsupervised model is trained on natural sentences only while the supervised model is trained on bilingual data which may have human or machine translations on the target side. Such discrepancies indicate that differences in system output may exist and prompt further investigation.

4.2 Quality Bins

Human evaluation complements automatic evaluation and allows abstraction away from the comparison to a human reference, which tends to favor the characteristics of translated text (Freitag et al., 2020). Professional linguists judge the output quality of Unsup and SupNC on a scale of 0–6. We divide the adequacy rating scale into bins of low (0–2), medium (3–4) and high (5–6) quality sentences. Table 3 reports the fraction of sentences judged to be in each quality bin, by system that produced it.

For subsequent analyses, it will be important to control both for quality and for content of output sentences so that neither become confounding variables. To compare linguistic aspects of alternative sentences with the same quality, we further analyze sentences falling into the same quality bin from both systems. For each source sentence, there is one translation by Unsup and one by SupNC. If human judges decide that both translations belong in the same quality bin, we add it to the “Both” row of Table 3. This results in 86, 255, and 218 sentences for low, medium, and high buckets, respectively.

4.3 Measuring Structural Proximity

To measure the influence of the source structure on the structure of the output, we develop a metric to account for changes in sentence structure without penalizing for differing word choice.

Word alignment appears well-suited to the task. Like Birch (2011), we calculate Kendall’s tau (Kendall, 1938) over alignments, but unlike them we do not simplify alignments into permutations. We calculate Kendall’s tau over fast_align (Dyer et al., 2013) alignments. We observe that the alignment algorithm struggled to align words not on the diagonal, meaning that alignments were sometimes skipped. This may make the correlation coefficient deceptively high.5

Alternatively, we propose translation edit rate (TER, Snover et al. (2006)) over part-of-speech tags as a novel research contribution that avoids this problem and is more suitable to the goal of measuring monotonic translation. TER is a well-known word-level translation quality metric which measures the number of edit operations required to transform an output sentence into the reference, and reports a “rate” by normalizing by sentence length. We compute TER over POS tags, so that TER now measures changes in structure independent of word choice. Source and target POS sequences which can be mapped onto each other with few edits are considered similar—a sign of a monotonic translation. Outputs with identical POS patterns score

5We ran fast_align with and without diagonal-favoring and all 5 symmetrization heuristics, and see similar trends.
Table 4: Perplexity of MT output on nt18 based on LMs trained on natural text vs. translated text, bucketed by quality. “Human” is the German reference side of nt18 (orig-en only). SupNC and Unsup are comparable supervised and unsupervised MT systems, respectively.

Table 5: TER (0-1+) over Universal Dependencies of human references vs. the source (nt18, orig-en).

We also compare the structure of alternative translations with finer-grained language-specific POS tags, i.e. German TIGER tags. Finer-grained labels avoid mapping all classes to the 17 UD categories valid across languages. For German this allows distinguishing sub-verb categories like infinitive, finite, model, and imperative (8 categories total). This monolingual POS comparison allows evaluation of whether systems’ output sentences have the same structure. We suspect that finer monolingual categories might uncover differences between systems that broad categories conceal. We run TER over German TIGER tags for all systems compared to the reference (row “All”) (Table 6). SupWMT14’s structure is most similar to the reference. SupNC and Unsup, systems with comparable BLEU, show negligible difference at this level of granularity, but subsequent rows bucketed by quality show the unsupervised output as less monotonic on the high-end of quality, while being more monotonic on the low-end. This finding suggests systematic difference in translation, and that unsupervised translations might be preferred as human-evaluated quality improves (See §4.2, Table 3).

4.4 Measuring Naturalness

Edunov et al. (2020) recommend augmenting BLEU-based evaluation with perplexity (ppl) from a language model (LM) to assess the fluency or naturalness of MT output. Ppl (Jelinek et al., 1977) measures how similar a text sample is to the model’s training data.

We contrast the likelihood of output according to two LMs: one trained on translated text (tLM) and another trained on non-translated natural text (nLM). While machine-translated text differs from human-translated text, the LMs are nonetheless a valuable heuristic and contribute insights on whether systematic differences between MT system outputs exist. Low ppl from the LM trained on natural German is an indication of natural language. Low ppl from the model trained on synthetic German (machine-translated English News Crawl) shows proximity to training data composed of translated text, indicating simplified language.

A model that distinguishes changes in naturalness of MT output should detect the same from human translation. Table 4 shows that human translation follows intuition, i.e. more likely according to the tLM than the nLM (26.31 vs. 43.82 ppl). SupNC ppl is lower than Unsup across quality bins for the tLM. Conversely, SupNC reports higher ppl
Table 7: SacreBLEU on nt18 of 4 MT systems, with perplexity (ppl) from LMs trained on natural or translated text. Lower ppl = model prefers the output. SupNC is supervised, trained on News Commentary. Sup En-Trns/De-Orig is trained on German-original News Crawl and synthetic English. Unsup is unsupervised, trained on natural English and natural German News Crawl. Unsup-Trns uses translated (synthetic) News Crawl only. Unsup is best according to both LMs, being more like natural text and less like translated text.

| System               | BLEU Overall | orig-en | orig-de | LM PPL Natural Text LM | Translated Text LM |
|----------------------|--------------|---------|---------|------------------------|--------------------|
| SupNC                | 29.2         | 34.0    | 21.1    | 72.97                  | 41.26              |
| Sup En-Trns/De-Orig  | 35.4         | 35.5    | 34.1    | 69.41                  | 50.40              |
| Unsup                | 30.1         | 30.9    | 27.1    | 66.69                  | 57.97              |
| Unsup-Trns           | 33.4         | 35.4    | 28.4    | 70.11                  | 48.91              |

by the nLM (except for high accuracy). All quality levels for Unsup have similar nLM ppl, suggesting it is particularly skilled at generating fluent output. These findings suggest that unsupervised MT output is more natural than supervised MT output.

5 Ablation: Architecture vs. Data

One reason that the unsupervised system might produce more natural-sounding output could be that it develops language-modeling capabilities from only natural German, whereas the supervised system sees synthetic data with the characteristics of translated text. In this section, we ask whether the improved naturalness and reduced monotonicity in unsupervised MT output is due to the different NMT architecture, or simply the data.

We build a supervised MT system using 329,000 paired lines of synthetic English source and natural German, where the source is back-translated German News Crawl from a supervised system. This allows the supervised system to also develop its language-modeling capabilities only on natural sentences. If more natural-sounding output is simply a response to training on natural German, then the supervised system should perform as well as the unsupervised system, or better.

Similarly, we train an unsupervised system on synthetic data only. The source-side of the training data is synthetic English from translating German News Crawl with a supervised system. The target-side is synthetic German which was machine-translated from English News Crawl. If language-modeling ppl is solely the result of changes in data, we expect this system to perform worst because it is trained only using synthetic data that exhibits some unnaturalness and monotonic translation.

Table 7 shows the results. We observe that the original unsupervised system (Unsup) performs best according to both LMs, having output that is more like natural text and less like translated text. When given only natural German from which to build a language model, the supervised system (Sup En-Trns/De-Orig) still produces output that appears more unnatural than Unsup. Even when the unsupervised system uses synthetic data only (Unsup-Trns), its output still appears more natural than the original supervised system (SupNC) according to both LMs. Taken together, these findings suggest that both German-original data and the unsupervised architecture encourage system output to be more natural-sounding, rather than data alone.

6 Leveraging Supervised and Unsupervised MT with BT

Our experimental results indicate that high-quality unsupervised output is less monotonic and more natural than supervised output. We are motivated to use the potential advantage to improve naturalness and decrease monotonicity in translation.

We explore ways to incorporate unsupervised MT into a supervised system via BT. For all experiments, we choose checkpoints based on validation performance on nt18, testing on nt19 and nt19p.

6.1 Baselines

We back-translate 24 million randomly-selected sentences of German News Crawl twice: once using a supervised German-English system trained on WMT18 bilingual training data with a transformer-big architecture, and once using our unsupervised system. Both use greedy decoding for efficiency. We augment the bilingual training examples of WMT18 (see §3) with either the supervised or unsupervised BT data to train two baselines. Table 8 shows the results. Supervised BT (+SupBT) performs as expected; minorly declining in BLEU on
the source-original test set (nt19 orig-en), improving on the target-original set (nt19 orig-de), and improving on the paraphrase set (nt19p). Conversely, adding unsupervised BT (+UnsupBT) severely lowers BLEU on source-original and paraphrase test sets. Randomly-partitioning the BT sentences such that 50% are supervised BT and 50% are unsupervised also lowers performance (+50-50BT).

6.2 Tagged BT

Following Caswell et al. (2019), we tag BT on the source-side. Tagging aids supervised BT (+SupBT_Tag) and greatly improves unsupervised BT (+UnsupBT_Tag), which outperforms the baseline and is nearly on-par with tagged supervised BT. Combining supervised and unsupervised BT using the same tag for both types (+50-50BT_Tag) shows no improvement over adding tagged supervised BT. We also try using different tags for supervised vs. unsupervised BT (+50-50BT_TagDiff). Decoding with tags during validation degraded performance across all conditions.\(^6\)

6.3 Probability-Based BT Selection

We design a BT selection method based on translation probability to exclude unsupervised BT of low quality. We assume that supervised BT is “good enough”. Given translations of the same source sentence (one supervised, one unsupervised) we assume that an unsupervised translation is “good enough” if its translation probability is similar or better than that of the supervised translation. If much lower, the unsupervised output may be low-quality. Selection runs as follows:

- Score each supervised and unsupervised BT with a supervised de-en system.
- Normalize the translation probabilities to control for translation difficulty and output length.
- Compare translation probability for each unsupervised and supervised BT of the same source sentence as:
  \[ \Delta P = \frac{P_{\text{norm(unsup)}}}{P_{\text{norm(sup)}}} \]
- Sort translation pairs by \(\Delta P\).
- Select the unsupervised BT for pairs scoring highest \(\Delta P\) and the supervised BT for the rest.

This is equivalent to filtering out unsupervised output sentences which are less than \(T\%\) as likely as the corresponding supervised sentence, where \(T\) is a hyperparameter, and swapping them with the corresponding supervised sentence. Importantly, our selection method results in the same 24M source sentences being used in all experiments. The selection procedure is shown in Figure 1.

The model we call “+MediumMix_Tag” uses the top \(\sim 40\%\) of ranked unsupervised BT with the rest supervised (9.4M unsupervised, 14.6M supervised), “+SmallMix_Tag” uses the top \(\sim 13\%\) of unsupervised BT (3.1M unsupervised, 20.9M supervised).\(^7\) We use the same tag for all BTs.

Table 8 shows the results. +SmallMix_Tag performs better than the previous best on nt18p and +MediumMix_Tag performs highest overall on nt19p. We recall that small differences on paraphrase test sets can signal tangible quality differences (Freitag et al., 2020). We trust the numbers on nt19p and use +MediumMix_Tag as our final model for human evaluation in the next subsection.

One might inquire whether improved performance is due to the simple addition of noise in light of Edunov et al. (2018), who conclude that noising BT improves MT quality. Subsequent work, however, found that benefit is not from the noise itself but rather that noise helps the system distinguish between parallel and synthetic data (Caswell et al., 2019).

\(^6\)Results were consistent whether model was selected by development performance on nt18 or nt18p.

\(^7\)The numbers are not round because data was selected using round numbers for the hyperparameter \(T\).
Joint (dev) | newstest2018 | orig-en | orig-de | nt18p | newstest2019 | orig-en | orig-de | nt19p  
--- | --- | --- | --- | --- | --- | --- | --- | ---  
Supervised Baseline (5.2M) | 41.8 | 46.1 | 34.3 | 12.6 | 38.8 | 30.4 | 11.7 |  
Unsupervised MT | 30.1 | 30.9 | 27.1 | 9.6 | 24.6 | 28.5 | 8.8 |  
**Supervised Baseline**  
+ SupBT | 43.4 | 43.7 | 41.8 | 12.5 | 37.0 | **39.9** | 12.0 |  
+ UnsupBT | 33.3 | 33.8 | 31.1 | 9.9 | 27.2 | 30.8 | 9.5 |  
+ 50-50BT | 38.0 | 36.4 | 39.0 | 12.9 | 29.4 | 38.3 | 10.0 |  
+ SupBT_Tag | **44.8** | 47.0 | 40.7 | 13.0 | **40.3** | 38.2 | 12.4 |  
+ UnsupBT_Tag | 43.3 | 46.9 | 36.9 | 12.9 | 39.1 | 35.0 | 12.2 |  
+ 50-50BT_Tag | 44.4 | **47.1** | 39.6 | 12.9 | 39.4 | 38.0 | 12.2 |  
+ 50-50BT_TagDiff | 44.4 | 46.8 | 40.1 | 13.0 | 39.9 | 37.9 | 12.4 |  
+ SmallMix_Tag | **44.8** | 46.8 | 40.8 | **13.2** | 39.8 | 38.8 | 12.5 |  
+ MediumMix_Tag | 44.7 | 46.8 | 40.8 | 13.0 | 40.1 | 38.2 | **12.6** |  

Table 8: SacreBLEU of a supervised baseline with 24M supervised or unsupervised back-translations. +MediumMix_Tag and +SmallMix_Tag utilize the BT selection method of §6.3. +MediumMix_Tag has 9.4M unsupervised BT and 14.6M supervised BT. +SmallMix_Tag has 3.1M and 20.9M, respectively. nt19p is the paraphrase reference set from Freitag et al. (2020), in which small BLEU score changes can indicate tangible quality difference.

---

2019; Marie et al., 2020). Yang et al. (2019) also propose tagging to distinguish synthetic data. With tagging instead of noising, Caswell et al. (2019) outperform Edunov et al. (2018) in 4 of 6 test sets for En-De, furthermore find that noising on top of tagging does not help. They conclude that “tagging and noising are not orthogonal signals but rather different means to the same end”. In light of this, our improved results are likely not due to increased noise but rather to systematic differences between supervised and unsupervised BT.

### 6.4 Human Evaluation

We run human evaluation with professional translators for our final model +MediumMix_Tag comparing its output translation of the nt19 test set with two of our baseline models. The quality evaluation in Table 9 shows that humans prefer the combined system over the baseline outputs.\(^8\) Table 10 shows the percentage of sentences judged as “worse than”, “about the same as”, or “better than” the corresponding +SupBT_Tag output, based on fluency. Raters again prefer the combined system.

### 7 Conclusion

We investigate whether unsupervised MT is useful in translation of high-resource language pairs. We performed the first systematic comparison of supervised and unsupervised MT output and propose a new metric for measuring monotonicity of translations. Our experiments indicate that unsupervised MT output is systematically different than supervised output, and our metrics point in the direction of increased naturalness of high-quality unsupervised output for English-German. We train an unsupervised back-translation augmented system that outperforms a traditional supervised system augmented with supervised back-translations on human-evaluated measures of quality and fluency.

---

\(^8\)Overall quality scores are low because the systems use only WMT18 bitext plus BT, and because professional translators typically score more harshly than crowd workers.

---

| Quality |  
| --- | --- | --- | --- | --- | --- | --- | --- |  
| + UnsupBT_Tag | 54.82 |  
| + SupBT_Tag | 56.13 |  
| + MediumMix_Tag | **58.62** |  

Table 9: Human-evaluated quality of supervised systems trained on WMT18 bitext +24 million backtranslated sentences, scored by professional translators.

| Rating |  
| --- | --- | --- | --- |  
| Worse | 45.2\% |  
| About The Same | 3.7\% |  
| Better | 51.1\% |  

Table 10: Side-by-side fluency eval. Shown: \% of +MediumMix_Tag sentences that professional translators judged “worse than”, “about the same”, or “better than” +SupBT_Tag output.
References

Stefanie Albert, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Sabine Brants, Thorsten Brants, Vera Dembger, Stefanie Dipper, Peter Eisenberg, et al. 2003. tigre annotationsschema. Universität des Saarlandes and Universität Stuttgart and Universität Potsdam.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2019. An effective approach to unsupervised machine translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 194–203, Florence, Italy. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In Proceedings of the Sixth International Conference on Learning Representations.

Mona Baker. 1993. Corpus Linguistics and Translation Studies: Implications and Applications. Text and technology: in honour of John Sinclair, pages 233–252.

Alexandra Birch. 2011. Reordering metrics for statistical machine translation.

Alexandra Birch, Phil Blunsom, and Miles Osborne. 2009. A quantitative analysis of reordering phenomena. In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 197–205.

Alexandra Birch and Miles Osborne. 2011. Reordering metrics for mt. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1027–1035.

Alexandra Birch, Miles Osborne, and Phil Blunsom. 2010. Metrics for mt evaluation: evaluating reordering. Machine Translation, 24(1):15–26.

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 745–754, Honolulu, Hawaii. Association for Computational Linguistics.

Arianna Bisazza and Marcello Federico. 2016. A survey of word reordering in statistical machine translation: Computational models and language phenomena. Computational linguistics, 42(2):163–205.

Nikolay Bogoychev and Rico Sennrich. 2019. Domain, Translationese and Noise in Synthetic Data for Neural Machine Translation. arXiv preprint arXiv:1911.03362.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (wmt18). In Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers, pages 272–307, Belgium, Brussels. Association for Computational Linguistics.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. Research on language and computation, 2(4):597–620.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers), pages 53–63, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc’Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 489–500.

Sergey Edunov, Myle Ott, Marc’Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 2836–2846.

Marina Fomicheva and Lucia Specia. 2016. Reference bias in monolingual machine translation evaluation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 77–82, Berlin, Germany. Association for Computational Linguistics.

Heidi Fox. 2002. Phrasal cohesion and statistical machine translation. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002), pages 304–3111. Association for Computational Linguistics.
Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at Scale and Its Implications on MT Evaluation Biases. In Proceedings of the Fourth Conference on Machine Translation, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Vireesh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be Guilty but References are not Innocent. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Online. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation.

Fred Jelinek, Robert L. Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America, 62(1/2):S63–S63.

Maurice G Kendall. 1938. A new measure of rank correlation. Biometrika, 30(1/2):81–93.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT ’11, pages 1318–1326, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In Proceedings of MT-Summit XII, pages 81–88.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. CoRR, abs/1804.07755.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Adapting translation models to translationese improves SMT. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL ’12, pages 255–265, Stroudsburg, PA, USA. 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. arXiv preprint arXiv:2001.08210.

Qingsong Ma, Yvette Graham, Timothy Baldwin, and Qun Liu. 2017. Further investigation into reference bias in monolingual evaluation of machine translation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2476–2485, Copenhagen, Denmark. Association for Computational Linguistics.

Benjamin Marie, Raphael Rubino, and Atsushi Fujita. 2020. Tagged back-translation revisited: Why does it really work? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 5990–5997, Online. Association for Computational Linguistics.

Anna Mauranen and Pekka Kujamäki. 2004. Translation universals: Do they exist?, volume 48. John Benjamins Publishing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in “multilingual” NMT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7737–7746, Online. Association for Computational Linguistics.

Larry Selinker. 1972. Interlanguage. International Review of Applied Linguistics, pages 209–241.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X. Chen, Ye Jia, Anjuli Kannan, Tara N. Sainath, and Yuan Cao et al. 2019. Lingvo: a Modular and Scalable Framework for Sequence-to-Sequence Modeling. CoRR, abs/1902.08295.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. 2020. Leveraging monolingual data with self-supervision for multilingual neural machine translation. arXiv preprint arXiv:2005.04816.
Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhou. 2006. A study of translation edit rate with targeted human annotation. In Proceedings of association for machine translation in the Americas, 6. Cambridge, MA.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In ICML.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In Proceedings of Machine Translation Summit XVII Volume 1: Research Track, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Antonio Toral. 2020. Reassessing Claims of Human Parity and Super-Human Performance in Machine Translation at WMT 2019. arXiv preprint arXiv:2005.05738.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In Proceedings of the Third Conference on Machine Translation: Research Papers, pages 113–123, Belgium, Brussels. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2203–2213.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In Proceedings of Machine Translation Summit XVII Volume 1: Research Track, pages 222–232, Dublin, Ireland. European Association for Machine Translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 5998–6008. Curran Associates, Inc.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 977–984, Sydney, Australia. Association for Computational Linguistics.