Like Chalk and Cheese?
On the Effects of Translationese in MT Training

Samuel Larkin Samuel.Larkin@nrc-cnrc.gc.ca
Michel Simard Michel.Simard@nrc-cnrc.gc.ca
Rebecca Knowles Rebecca.Knowles@nrc-cnrc.gc.ca
National Research Council Canada, Ottawa, Ontario, Canada

Abstract

We revisit the topic of translation direction in the data used for training neural machine translation systems, focusing on a real-world scenario with known translation direction and imbalances in translation direction: the Canadian Hansard. According to automatic metrics, we observe that using parallel data that was produced in the "matching" translation direction (authentic source, translationese target) improves translation quality. In cases of data imbalance in terms of translation direction, we find that tagging the translation direction of training data can close the performance gap. We perform a human evaluation that differs slightly from the automatic metrics, but nevertheless confirms that for this French–English dataset that is known to contain high-quality translations, authentic or tagged mixed source improves over translationese source for training.

1 Introduction

Prior work in statistical machine translation (SMT) highlighted potential benefits of making use of information about the translation direction of training data (Kurokawa et al., 2009). When text is translated, there is an *authentic source* (the language in which the text was originally produced), and its translation, which in contrast can be described as *translationese*. Thus when considering translation direction in machine translation, training data can be described as consisting of *authentic source*, *translationese source*, or a mix.[1] Backtranslated data produced by machine translation may be thought of as an extreme case of translationese source (Marie et al., 2020), but because the quality and types of errors that occur in machine translation are quite different from those that occur in human translation, it is worth examining translation direction of human translation separately from MT-based data augmentation. In Figure 1 we show a fairly dramatic example of the kinds of translation quality differences that can occur when building MT systems using authentic source as opposed to translationese source.

Recent work in neural machine translation (NMT) has revisited this issue, motivating the automatic detection of (human) translationese by showing improved performance on several metrics when training translation direction matches the testing translation direction (Sominsky and Wintner, 2019), examining domain and backtranslation along with the translation direction of test sets (Bogoychev and Sennrich, 2019), and evaluating the treatment of predicted translation direction as separate languages in a multilingual-style NMT system through human and automatic metrics (Riley et al., 2020).

[1] For the purposes of this paper, we will set aside the situation where both sides of the text consist of translationese, translated from one or more other pivot languages.

Source	Les producteurs de fromage au Québec sont des fleurons dont on est fiers.
Reference	We are proud of our exceptional Quebec cheese producers.
MT (Authentic Src.)	We are proud of the success of cheese producers in Quebec.
MT (Translationese Src.)	The cheese producers in Quebec are proud flowers.

Figure 1: Example output of French-English MT trained on Authentic-source (authentic French, translationese English) and Translationese-source (translationese French, authentic English).

We focus this work on a particular real-world scenario, where translation direction is known, and translation (whether human, machine, or computer aided) is expected to be performed from authentic source language text. This is, in fact, a fairly common scenario (i.e., parliamentary, legal, medical, patent, etc. translation), and we highlight one such case as an example: the Canadian Hansard (House of Commons), which consists of transcripts of parliamentary speech, alongside their translations. These proceedings are published in French and English, and it is indicated whether the authentic source was French or English.[2] There is also an imbalance in translation direction; most of the text of the Hansard was originally spoken in English and transcribed and then translated into French. Given that the text is formal and falls within the parliamentary domain, it is appropriate to build or adapt translation systems using the existing Hansard as training data, for use in translating future Hansard text (i.e., in a computer aided translation setting), which raises questions about how to make the best use of the available text and the metadata regarding source language.

In this work, we focus on translating original (authentic) source language text. We examine the following questions:

Q1: What effect does translation direction of training data have on system output?

Q2: Can tagging source side translationese in the training data (i.e., adding a special token like "<translationese>" to the start of translationese source sentences) improve translation of authentic source language test data?

Q3: In a moderate resource setting (approx. 3.7 million sentence pairs), what effect does the proportion of source side translationese (from 0% to 100%) in the training data have?

We experiment and evaluate these using automatic metrics and a small human judgment task, looking at both French–English and English–French translation directions. With regard to **Q1**, we find that systems trained exclusively with Authentic source data outperform by a large margin those trained exclusively with Translationese source data, even with twice as much training data. Combining Authentic and Translationese source does not always produce significantly better systems, compared to using Authentic source only, but tagging Translationese source in the training and tuning data (**Q2**) can improve performance, especially in situations where there is more Translationese source than Authentic source data. In general, translation quality increases as the percentage of Authentic source training data increases (**Q3**): below 50%, tagging Translation source data can help bridge the gap, but the importance of tagging decreases as the percentage of Authentic source training data increases.

2 Data

We use parallel English–French (EN-FR) text from the Canadian Hansard, House of Commons. Our corpus contains transcripts of debates from 1986 to 2016. Earlier parts of this dataset are

[2] Other languages are spoken in the House of Commons, notably Indigenous languages, but in those cases, English and French translations are provided in the Hansard. (`https://www.ourcommons.ca/DocumentViewer/en/42-1/PROC/report-66/`)

available from LDC (Ma, 1999), more recent transcripts are publicly available from the Canadian Parliament website.[3] This data is known to have high translation quality. It is annotated with direction of translation (the original language, FR or EN, as spoken in the House is known); we omit all lines marked as unspecified.[4]

The question of domain is always intertwined with the question of translation direction. Here we hope to minimize that by confining our work to the parliamentary domain; we expect that the level of formality and style of parliamentary speech is relatively consistent, even across languages (certainly more so than it would be if compared between news data and parliamentary speech). Nevertheless, we acknowledge that there will remain differences within this domain; i.e., Members of Parliament may speak more frequently about topics related to their own constituencies or about different topics over time. We also sample our data with an eye toward temporal aspects for this reason.

The full dataset (from which we select our training, development, and test data) is unbalanced in terms of original language: 10,091,250 lines (68.5%) were originally spoken in English, while 3,699,822 lines (25.1%) were originally spoken in French (the remaining 933,996 lines, 6.3%, were labeled as unspecified). In order to run experiments on the proportion of source-side translationese used, we are limited by the size of the smaller sub-corpus, the Authentic-FR language data.

We sample data for validation and testing (2k and 8k lines, respectively), with Authentic-EN source data used for translation into FR and vice versa. The validation and testing data are randomly sampled sentences from recent data (Nov. 1 to Dec. 15, 2016), while training contains older data. This mimics a real-world scenario, where translators (potentially using computer aided translation) might post-edit or interactively translate new text using the output of machine translation systems build on older text. By drawing the test sentences from a separate portion of the Hansard as the training data, we guarantee that test sentence performance is not inflated due to having included neighboring context in the training data; rather, the test data performance should be representative of realistic performance on new and previously unseen Hansard data.

For Q3, we subsample Authentic-EN parallel text once, to match the Authentic-FR training data in size, also attempting to match it in date distribution (which we expect may also serve as a proxy for matching topic distributions).[5] For the experiments that consider between 0% and 100% source side Translationese, we then subsample this Authentic-EN subsample and the Authentic-FR data.

We preprocess the data using open-source normalization and tokenization scripts from `PortageTextProcessing`.[6] Specifically, we applied `clean-utf8-text.pl` (removing control characters, standardization, etc.), followed by `fix-slashes.pl` (heuristically adding whitespace around slashes), and tokenization with `utokenize.pl -noss -lang=$lang`. We then train joint 32k byte-pair encoding (BPE) subword vocabularies on the training data (Sennrich et al., 2016),[7] and apply them to train, development, and test.

3 Models

We build Transformer (Vaswani et al., 2017) models using Sockeye-1.18.115 (Hieber et al., 2018), with 6 layers, 8 attention heads, network size of 512 units, and feedforward size of 2048 units. We have changed the default gradient clipping type to *absolute*, used source-target soft-

[3] `https://www.ourcommons.ca/`

[4] This includes both boilerplate text and full sentences.

[5] We read in the corpus chronologically, maintaining counts of Authentic-EN and FR and sampling from a Gaussian to determine whether to keep or discard incoming original-EN sentences to maintain similar counts.

[6] `https://github.com/nrc-cnrc/PortageTextProcessing`

[7] `https://github.com/rsennrich/subword-nmt`

max weight tying, an initial learning rate of 0.0002, batches of $\sim$8192 tokens/words, maximum sentence length of 200 tokens, optimizing for BLEU, checkpoint intervals of 4000, and early stopping after 32 checkpoints without improvement. Decoding used beam size 5. Training used 4 NVIDIA V100 GPUs.

4 Experiments

4.1 Challenges and Evaluation

We measure system quality through automatic metrics: BLEU (Papineni et al., 2002) and chrF (Popović, 2015), both of which we computed using SacreBLEU (Post, 2018). We show BLEU score 95% confidence intervals using bootstrap resampling (Koehn, 2004) with 1000 iterations of sampling the full test set (with replacement). When the confidence intervals are non-overlapping, we can claim statistically significant differences between the systems, but when they overlap we cannot directly make claims about statistical significance or the lack thereof. We also perform *pairwise* bootstrap resampling, again with 1000 iterations, in order to evaluate whether improvements from one system to another are statistically significant (Koehn, 2004). Recent work has noted that BLEU score can effectively be gamed by producing more translationese-like text (Riley et al., 2020), improving automatic metric scores while decreasing quality according to human ratings. Mathur et al. (2020) observe that small improvements in metric scores may not always result in corresponding improvements in human judgments. We address this by complementing BLEU with another metric (chrF) and doing a manual (human) analysis of translation quality.

For the human evaluation, we asked annotators to perform two sets of three-way ranking tasks on a sample of test sentences produced by three different systems. We then computed average rankings of the three systems based on the human judgments.[8] Annotators viewed a source sentence, its reference translation, and were asked to rank three translations of it based on which output they found to be the best translation (semantically, grammatically, and fluency-wise).[9] There was also a free text box for optional comments. The ranking was performed using LimeSurvey,[10] and the three sentences were displayed in a random order. All annotators first completed 100 annotations for interannotator agreement; we expected this to be quite low.

We measured interannotator agreement using Cohen's kappa coefficient (κ), as in Bojar et al. (2013):

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

where P(A) is the proportion of times that pairs of annotators agree on the relative ranking of pairs of systems, and P(E) is the proportion of times that they would agree by chance.[11] We find overall $\kappa = 0.25$ for EN-FR translations and $\kappa = 0.28$ for FR-EN. Such values of κ are typically interpreted as indicating "fair" agreement (Landis and Koch, 1977). If we convert the rankings into the task of labeling the *best* system, annotator agreement increases: $\kappa = 0.31$ (EN-FR) and $\kappa = 0.31$ (FR-EN). The agreement on which is the *worst* system is even stronger: $\kappa = 0.34$

[8] Annotators were adult L1/fluent speakers of the target language with knowledge (ranging from conversational to fluent) of the source language, including the authors and colleagues, five for French, four for English; all volunteer. No personally identifying information was collected.

[9] Annotators were only asked to judge sentence tuples where there were at least two unique translations of the sentence; exact matches ranked consecutively were scored as ties (such that the final ranking could be either: 1-2-3, 1-1-2, or 1-2-2). This explains why average ranks don't always sum to 6, as would be expected if all ranks were exclusive. 21 annotations where exact matches were ranked non-consecutively were dropped (out of a total of 1800 annotations, this is approximately 1%).

[10] https://www.limesurvey.org/

[11] κ is calculated excluding comparison of identical system outputs.

(EN-FR) and $\kappa = 0.39$ (FR-EN). The example sentence pair in Figure 1 is an extreme one, showing the worst effects of translationese training. In their qualitative assessments, annotators noted that this was a challenging ranking task, as the sentences they were judging often differed by only a few tokens; several annotators expressed a wish for a mechanism for marking ties. In many cases this was an issue of three high-quality outputs, though there were also examples of three equally-poor outputs.

As with BLEU scores, we compute 95% confidence intervals around the average rankings using bootstrap resampling of the human ranking data (Koehn, 2004) with 1000 iterations of sampling the full annotated sets with replacement. We also perform pairwise bootstrap resampling for significance.

In the following sections, we discuss both the automatic and the human rankings in greater detail, including the matter of statistical significance (via confidence intervals and pairwise bootstrap resampling), where the human and automatic metrics agree and disagree, and what trends we observe that do not rise to the level of statistical significance but which may still merit future work.

4.2 Q1: Translation Direction

We first examine the effects of translation direction in our realistic setting, considering three systems built with three different training sets: Authentic source only, Translationese source only, and finally their combination (Mixed; all available data). As we evaluate by translating Authentic source data, we expect that training on Authentic source data should be better than training on Translationese source data.

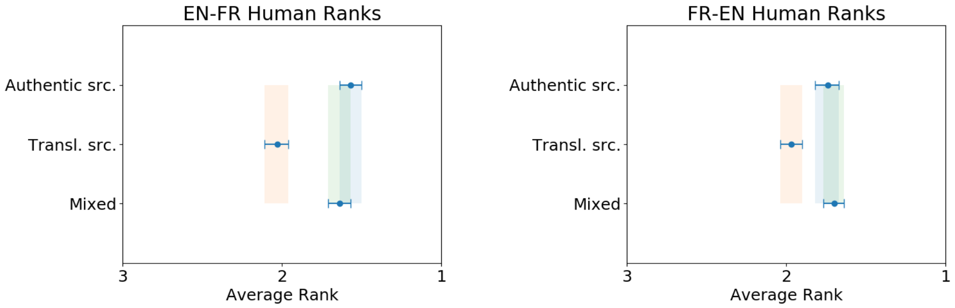
	EN→FR				FR→EN			
	Lines	BLEU ↑	chrF ↑	Human ↓	Lines	BLEU	chrF	Human
Auth. Src.	10.0M	42.8 ± 0.6	0.651	1.57	3.7M	52.0 ± 0.7	0.716	1.74
Transl. S.	3.7M	38.0 ± 0.6	0.616	2.03	10.0M	48.0 ± 0.7	0.689	1.97
Mixed	13.7M	43.0 ± 0.6	0.652	1.64	13.7M	52.0 ± 0.7	0.715	1.70

Table 1: Comparison of translation quality of systems trained on Authentic source only, Translationese source only, or the combination of the two, measured in terms of BLEU (with 95% confidence intervals) and chrF on the test data. The *Human* column reports the average ranking of the system (1 is the best, 3 is the worst). The *Lines* column shows the number of lines used in training the system.

Table 1 shows the results. As expected, in both translation directions, using Authentic source data for training outperforms using Translationese source data (by a difference of 4.8 BLEU in the EN-FR direction and by a difference of 4.0 in the FR-EN direction). This is particularly striking in the FR-EN direction: despite using more than twice as much training data (10.0M lines as compared to 3.7M), the Translationese source condition lags well behind the Authentic source condition by all metrics. We conclude that the Translationese source system is significantly worse than the Authentic source and Mixed source systems, as evidenced by the non-overlapping 95% confidence intervals and the fact that 100% of pairwise bootstrap resampling iterations found the Translationese to be worse than either system it was paired with.

The performance of training with the Mixed data is very comparable to training with only Authentic data. In the EN-FR direction, there is a difference of 0.2 BLEU in favor of the Mixed training data, while in the FR-EN direction there is a very small difference of 0.08 BLEU. According to pairwise bootstrap resampling of BLEU scores, the EN-FR Mixed system is significantly better than the Authentic only system ($p < 0.05$, with the Mixed system performing better in 95.7% of resampling instances). In the FR-EN direction, the BLEU difference between

Mixed and Authentic is not statistically significant. In the EN-FR direction, chrF also shows a small gain for the Mixed training data, while in the FR-EN direction, the Authentic source has a very small advantage.



(a) English–French

(b) French–English

Figure 2: Confidence intervals (shaded for visibility) for average human-annotated rank (rank 3 is worst and rank 1 is best) for systems corresponding to Table 1.

We turn to human evaluation, where for systems in Table 1, we have 395 (EN-FR) and 498 (FR-EN) annotations respectively. We found that the human rankings agreed with the automatic metrics in terms of which system was consistently worst: the Translationese source system. As evidenced by the distinctly non-overlapping 95% confidence intervals in Figure 2 and via pairwise bootstrap resampling with $p = 0.05$, human judgments (like automatic metrics) judge the Translationese source model to be significantly worse than each of the other two. This result contrasts with Riley et al. (2020).

Annotators disagreed slightly with automatic metrics in terms of ranking Authentic source and Mixed source, but we note that the differences between those scores (both automatic and human) were quite small. For EN-FR, automatic scores scored the Mixed source best by 0.2 BLEU and 0.001 chrF, while human judgments scored Authentic source systems as best by an average rank difference of 0.07. For FR-EN, BLEU had Authentic and Mixed source tied, while chrF had Authentic source edging out Mixed by a difference of 0.001; human rankings favored the Mixed by 0.04. While these results are not *statistically* significant (for human rankings), they do raise questions about the effects of the ratio of Authentic and Translationese source data, which we examine in more detail in Section 4.4.

When testing the above systems on Translationese source test data, we observe results similar to the ones discussed here: in that setting, systems trained on Translationese source perform better than systems trained on Authentic or Mixed data. However, since our primary interest is in the more realistic task setting of translating Authentic source data, we do not further discuss these results here.

We note that the data is unbalanced, with much more Authentic English source than Authentic French source, due to the distribution of language as spoken in the House of Commons. The fact that using Authentic source training data performs better when translating Authentic source test data than Translationese source data (even when there is *much* more Translationese source data) indicates that translation direction does matter.

4.3 Q2: Tagging Translation Direction

Having observed through automatic and human metrics that the translation direction does matter, we turn to the question of tagging translation direction (with a special “<translationese>” tag at the start of the source sentence for source-Translationese sentences), to see if this will enable the Mixed data systems to make better use of all available information. Tagging has been shown to be effective in multilingual (Johnson et al., 2017; Rikters et al., 2018) and multi-domain (Kobus et al., 2017) systems, as well as when using backtranslated data (Caswell et al., 2019). All of these systems for Q2 make use of the full 13.7M line training set.

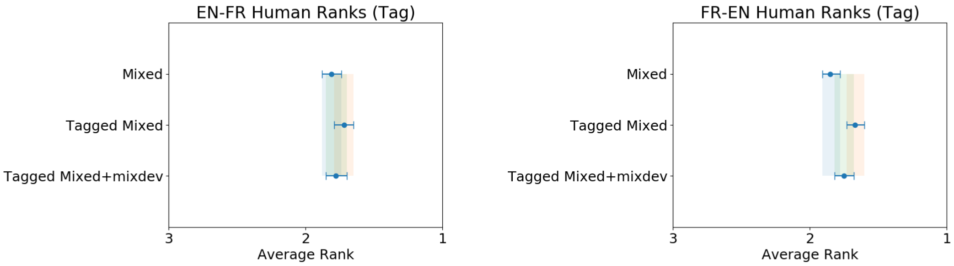
	EN→FR			FR→EN		
	BLEU ↑	chrF ↑	Human ↓	BLEU	chrF	Human
Mixed	43.0 ±0.6	0.652	1.81	52.0 ±0.7	0.715	1.85
Tagged Mixed	43.0 ±0.6	0.653	1.72	52.6 ±0.7	0.720	1.67
Tagged Mixed+mixdev	43.1 ±0.6	0.653	1.78	52.9 ±0.7	0.722	1.75

Table 2: BLEU and chrF scores for training with Mixed data, untagged and tagged (the latter with Authentic source validation or Mixed validation). We indicate if a system is tagged through the addition of "Tagged" in the system name, while untagged systems are unmarked. The untagged systems here are the same Mixed systems shown in Table 1.

Table 2 shows our results. The effect of tagging is stronger in the FR-EN translation direction, where simply adding tags results in a BLEU score increase of 0.6 (chrF increase of 0.005). We recall that the Authentic FR source data is much smaller than the Authentic EN source, so we hypothesize that tagging allows the system to take better advantage of the two types of data. In the other translation direction, where Authentic EN source already comprises the majority of the training data, we observe minimal changes when applying tagging.

The Mixed (untagged) and initial Tagged Mixed experiments are performed with a validation set that consists only of Authentic source data. This raises the question of whether that is adequate to make the most of the information contained in the tags, or whether using a Tagged Mixed validation set (with 1461 lines Authentic-EN source, and 539 lines Authentic-FR source) might be better. We refer to this system that uses the Tagged Mixed validation set as "Tagged Mixed+mixdev" in Table 2. In the FR-EN direction, we see an additional 0.3 BLEU improvement when using the Tagged Mixed+mixdev (0.002 chrF improvement). In the other direction, we see a small 0.1 BLEU improvement and no corresponding change in chrF. In the EN-FR direction, where Authentic data was already the majority, we do not find any significant BLEU score differences between the various tagged and untagged systems. However, in the FR-EN direction, both the Tagged Mixed and Tagged Mixed+mixdev systems are found to be significantly better in terms of BLEU than the Mixed (untagged) system, according to pairwise bootstrap resampling (with 100% of samples showing this to be the case). Paired bootstrap resampling also finds that in the FR-EN direction the Tagged Mixed+mixdev system is significantly better in terms of BLEU than the Tagged Mixed system ($p < 0.05$, with 98.6% of the samples showing this result).

Human evaluation provides additional insight. For Table 2 systems, we collected rankings for 391 (EN-FR) and 495 (FR-EN) source sentences and their translation triplets, respectively. We first note that in both translation systems, we observe the same pattern: Tagged Mixed is ranked best, followed by Tagged Mixed+mixdev, with Mixed (untagged) ranked worst. In the English–French direction, none of the average human system rankings differ significantly, which is unsurprising given how close they are to one another, as shown in Figure 3a. This matches the automatic metrics and our intuitions: Authentic English source makes up the majority of the Mixed training data, and we already observed that Authentic and Mixed translation systems performed quite similarly in this direction. In the French–English direction, shown in Figure 3b, we do not find a significant difference in human rankings between the the two tagged systems (Tagged Mixed and Tagged Mixed+mixdev). However, based on pairwise bootstrap resampling, the human annotators rank both tagged systems (Tagged Mixed and Tagged



(a) English–French

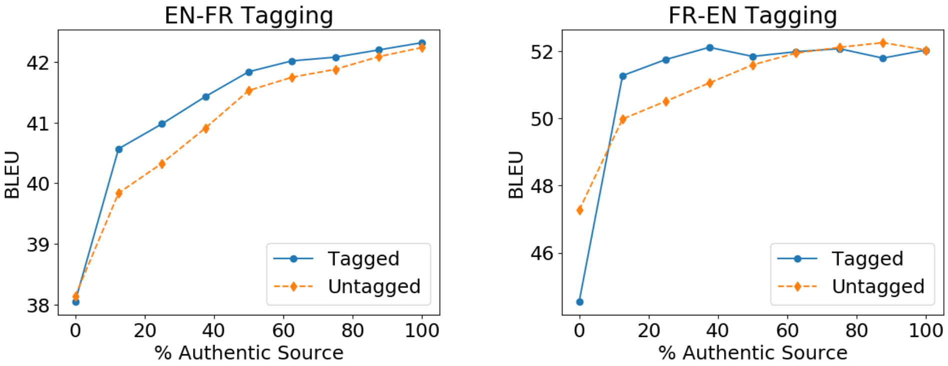
(b) French–English

Figure 3: Confidence intervals (shaded for visibility) for average human-annotated rank (rank 3 is worst and rank 1 is best) for systems corresponding to Table 2 (effects of tagging).

Mixed+mixdev) significantly higher than the (untagged) Mixed system. This is partially in agreement with the results on BLEU, but may merit more exploration.

The significant improvement in human ranking by adding tagging (FR-EN) suggests that in a scenario where the Authentic source data makes up a minority of the training data, it is beneficial to add direction tags. When Authentic data makes up the majority of the training data, it does not *hurt* to add direction tags, but it does not appear to significantly help. We examine this in a controlled experiment in Section 4.4.

4.4 Q3: Proportion of Source Translationese



(a) English–French

(b) French–English

Figure 4: Effect of tagging and percentage of Authentic source data.

In our previous experiments, we maintained a fixed ratio of Authentic and source Translationese, matching the true distribution of our dataset. We now examine what happens when we vary the ratio of Authentic to Translationese source data, maintaining a fixed corpus size. This is a moderate resource setting with 3.7M lines.[12] We vary the proportion of Authentic training data from 0% to 100% (by steps of 12.5%) and build translation systems in both directions, both tagged and untagged, using Authentic source validation sets. As we see in Figure 4, translation quality on Authentic source test data increases as the percentage of Authentic source training

[12] As described in Section 2, this consists of all Authentic French source and a sample of Authentic English source, subsampled to vary proportions.

data increases. Below 50%, tagging clearly helps bridge the gap, but the importance of tagging decreases as the percentage of Authentic source training data increases. This trend matches our earlier intuitions. In the English–French direction (Figure 4a), the gap is greatest at 12.5% (i.e., tagging provides the most additional benefit), and shrinks as it approaches in 100%. In the French–English direction (Figure 4b), the story is similar, though the two approaches appear to converge around 50%.[13] Thus we would argue that tagging translation direction is worth considering in situations where the "matching" translation direction (Authentic source) makes up the minority of the data, though it may still have some benefits at higher percentages.

5 Conclusion

We have shown that in a moderate-resource setting with high-quality translations in training data, training on Authentic-source data or Tagged Mixed-source data is preferred over training on Translationese-source or Mixed (untagged) source data, by both automatic metrics and human judgments. This is in contrast with the findings of Riley et al. (2020), who found that BLEU scores could be "gamed" to produce higher scores with translationese-like output, while being judged to be worse by human annotators. This raises questions for future work, such as whether Translationese training effects may vary depending on the quality of the parallel text, the proportion of Translationese data, and the size of the training data, or whether differences in experimental setup and human annotation may also come into play. Future work could examine these issues across a wider range of language pairs and domains, as well as directly comparing known translation direction with automatically predicted translation direction.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions and Gabriel Bernier-Colborne and Cyril Goutte for feedback and discussion. We thank the volunteer annotators for their time and assistance. This work was done as part of a collaboration with the Canadian Translation Bureau and was funded in part by Public Services and Procurement Canada.

References

Bogoychev, N. and Sennrich, R. (2019). Domain, translationese and noise in synthetic data for neural machine translation. *CoRR*, abs/1911.03362.

Bojar, O., Buck, C., Callison-Burch, C., Federmann, C., Haddow, B., Koehn, P., Monz, C., Post, M., Soricut, R., and Specia, L. (2013). Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria. Association for Computational Linguistics.

Caswell, I., Chelba, C., and Grangier, D. (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.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2018). The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

[13] We do note, in the French–English direction, that the tagged 0% system performs surprisingly poorly; we expect this is due to chance initialization issues, rather than anything specifically to do with the tags, which *should* have no effect in that scenario.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kobus, C., Crego, J., and Senellart, J. (2017). Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *In Proceedings of MT-Summit XII*, pages 81–88.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ma, X. (1999). Parallel text collections at linguistic data consortium. In *Machine Translation Summit VII, Singapore*.

Marie, B., Rubino, R., and Fujita, A. (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.

Mathur, N., Baldwin, T., and Cohn, T. (2020). Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Popović, M. (2015). chrF: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Rikters, M., Pinnis, M., and Krišlauks, R. (2018). Training and adapting multilingual NMT for less-resourced and morphologically rich languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan. European Languages Resources Association (ELRA).

Riley, P., Caswell, I., Freitag, M., and Grangier, D. (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.

Sennrich, R., Haddow, B., and Birch, A. (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.

Sominsky, I. and Wintner, S. (2019). Automatic detection of translation direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.