Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6759–6767
Marseille, 20-25 June 2022

Translation Memories as Baselines for Low-Resource Machine Translation

Rebecca Knowles, Patrick Littell
National Research Council Canada
1200 Montreal Road, Ottawa ON, K1A 0R6
{Rebecca.Knowles, Patrick.Littell}@nrc-ncrc.gc.ca

Abstract

Low-resource machine translation research often requires building baselines to benchmark estimates of progress in translation quality. Neural and statistical phrase-based systems are often used with out-of-the-box settings to build these initial baselines before analyzing more sophisticated approaches, implicitly comparing the first machine translation system to the absence of *any* translation assistance. We argue that this approach overlooks a basic resource: if you have parallel text, you have a translation memory. In this work, we show that using available text as a translation memory baseline against which to compare machine translation systems is simple, effective, and can shed light on additional translation challenges.

Keywords: machine translation, low-resource, translation memories

1. Introduction

The goal of machine translation (MT) is to take as input heretofore unobserved sequences of text in the source language and produce as output accurate and fluent translations of that same text in the target language. When it comes to low-resource language pairs, this is an especially challenging task. When building supervised[1] machine translation systems or introducing a new dataset for a low-resource language pair, the typical approach is to build baseline systems using well-known neural or phrase based statistical machine translation architectures using all available parallel text in the languages (perhaps aside from a held-out test and/or validation set). For some language pairs, this could be on the order of just thousands of sentence pairs. While the field continues to progress, automatic and human evaluations of many low-resource machine translation systems demonstrate that they are still far from reaching the point where they can be consistently useful for machine translation's downstream applications: comprehension, communication, or publication.

Very low-resource machine translation is often of too low quality to use directly, without any sort of intervention or improvement. Consequently, researchers might hope that an initial machine translation system could be useful to human translators as part of a computer aided translation (CAT) pipeline (i.e., post-editing, interactive translation, etc.). When we build a first machine translation baseline, that first baseline is implicitly compared against having *no* translation assistance at all. However, many translators working in CAT workflows do have access to translation tools: dictionaries, spell checkers, and *translation memories*. A translation memory (TM) is a collection of source-target segment pairs which are (human-produced and/or human-validated) translations of one another. When preparing to translate a novel sentence, the CAT tool can present translations of similar sentences from the TM to the translator, which they can then modify.

If you have parallel text suitable for machine translation, you have a translation memory. (Though of course this may not be a *well-curated* translation memory; it may be noisy, or may contain text that translators would not have chosen to keep in a translation memory, etc.) In order to see how much a machine translation system might contribute to a computer aided translation pipeline, comparing a machine translation system to a *translation memory* baseline is a fairer comparison than comparing it against no assistance. As such, we propose that researchers compare against a TM baseline, in addition to any desired machine translation baselines and as a complement to human evaluation and analysis. At the AmericasNLP Shared Task (Mager et al., 2021), TM baselines proved to be strong baselines across several languages, even outperforming (according to automatic metrics) some trained systems that incorporated additional data, but falling short of the state of the art systems. We also demonstrate that a simple TM baseline can provide information about similarities and differences between available training and testing data, which can be useful in determining appropriate algorithms and preprocessing. We argue that examining datasets through a translation memory lens can provide a way of categorizing different types of low-resource tasks. Finally, we take a fine grained look at how TM baselines can help us conceptualize the potential usefulness (or lack thereof) of a given machine translation system on a per-sentence basis, showing how translation systems may succeed or fail to generalize well from the available data. We demonstrate this using data across a range of recent low-resource language translation tasks.

[1] In this work, we focus only on supervised machine translation, not the scenario where there is *no* parallel text available.

Langs.	Train	Dev.	Test
hch-es	9k	994	1003
tar-es	15k	995	1003
nah-es	16k	672	995
gn-es	26k	995	1003
hsb-de	147k	2000	2000
iu-en	1300k	5173	2971

Table 1: Lines of training (rounded to the nearest 1000), development, and test data for language pairs.

2. Data

We explore the idea of TM baselines across datasets from several shared tasks on low-resource machine translation. We use the language codes corresponding to the files provided by the various datasets: de German, hsb Upper Sorbian, en English, iu Inuktitut, hch Wixárika, tar Rarámuri, nah Nahuatl, gn Guaraní, es Spanish.

The AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas (Mager et al., 2021) included translation from Spanish into 10 Indigenous languages of the Americas. The training and test data for some of these languages differed quite noticeably, including some situations with different dialects or spelling conventions. Combined with the very small data sizes, it is unsurprising that resulting human evaluations of even the best systems found their output to be mostly inadequate (with mixed fluency results). We examine a subset of four of the language pairs in this work.

Translation between Upper Sorbian and German was part of the Very Low Resource Supervised MT track at both WMT 2020 and WMT 2021 (Fraser, 2020; Libovický and Fraser, 2021). In contrast to the AmericasNLP shared task, the train and test data for these tasks were very closely matched, resulting in exceptionally high automatic metric scores for a low-resource task. We use the 2021 training data, development data, and devel_test (which we refer to here as test) for this work. Unfortunately, no human evaluations are available for the systems produced for these tasks.

Inuktitut–English MT was included in the News Translation shared task at the Fifth Conference on Machine Translation (WMT 2020, Barrault et al. (2020)). The parallel training data available for this task consisted mainly of sentence pairs from the Nunavut Hansard 3.0 data release of text from the Proceedings of the Legislative Assembly of Nunavut (Joanis et al., 2020). The test data was half parliamentary text (from recent sessions that were not included in the earlier data release) and half news data, from Nunatsiaq News (used with permission). There was also development/validation data provided from both domains. Human evaluation data was collected for both translation directions.[2]

While all of these tasks were described as "low resource" tasks, they are in fact quite varied in data size. Table 1 shows how the training data range in size from just under 9000 to approximately 1.3 million lines, while the development data are between 672 and 5173 lines, and the test data range from 995 lines to 2971. The values are listed as described in the papers describing the datasets (Mager et al., 2021; Libovický and Fraser, 2021; Joanis et al., 2020), with test sets measured separately.

What we have is a huge range of what may count as "low resource" for machine translation. Due to differences in language typology, it is not necessarily as simple as looking only at number of lines of training data; one may also wish to consider morphological complexity or other linguistic features. For example, Inuktitut is known to be highly morphologically complex, resulting in many words (defined as space/punctuation-separated) that appear just once or only a few times, even in such a large corpus. There is also the question of dialect, orthography, and domain matches. Knowles et al. (2021) compare these three tasks, and place them in a 2-by-2 matrix based on "match" vs. "mismatch" and "low-resource" vs. "mid-resource", which we reproduce here in Table 2.

	Domain Match	**Mismatch**
Low-Res.	Upper Sorbian	AmericasNLP
Mid-Res.	Inuktitut Hansard	Inuktitut News

Table 2: Comparison of three recent shared tasks on low-resource machine translation.

In this work, we expand upon these distinctions and their consequences for low resource machine translation research. ∀ et al. (2020) point out that there are many aspects to low-resourcedness beyond data. This work is not a substitute for the participatory research that they propose, but provides a complementary way to examine different kinds of low-resourcedness within data, which can then be explored in-depth with language experts.

3. TM Baselines

Given a new sentence to translate, a CAT tool that incorporates a TM typically uses a "fuzzy match" score to find the most similar source segment in the TM, along with its translation. We follow Simard and Fujita (2012) in using MT evaluation metrics as simple fuzzy match scores in our translation memory process. We use BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) as implemented in sacrebleu (Post, 2018) to compute the fuzzy match scores and to evaluate the resulting data.[3] While there may be benefits to using trained (i.e., embedding-based) metrics in many cases,

[2] However, Knowles (2021) notes that the English–Inuktitut human system ranking table in Barrault et al. (2020) may not reflect the final and complete data collection.

[3] For scores over full test sets, signatures are as: nrefs:1,case:mixed,eff:no,tok:13a,smooth:exp,version:2.0.0 (BLEU) and nrefs:1,case:mixed,eff:yes,nc:6,nw:0,space:no, version:2.0.0 (CHRF)

in the low-resource scenarios that we are examining, there is a risk of introducing confounds due to the limited data available to train the embeddings (and the likelihood of overlap with the data used for machine translation), so we use just these two metrics.

3.1. TM Definitions and Experimental Setup

Given an input sentence s in our test set (with reference translation t), we find the source sentence s' in our TM that maximizes $\mathcal{M}(s', s)$ where $\mathcal{M}$ is a function that measures similarity between two strings (in practice, BLEU or CHRF). For MT evaluation metrics that are not symmetric, like BLEU, where the length of the reference is used in computation, we use s as the "reference" and treat s' as a "hypothesis" to maintain consistency. We then return t' (the translation of s') as a hypothesis translation of s. We can then compute our MT evaluation metric of choice, comparing each reference t to its corresponding hypothesis t' (over the full development or test set). When computing metric scores for evaluation over the full development or test set, we use the corpus-level versions of the metrics. When computing metric similarity scores for $\mathcal{M}(s', s)$ we do this on a single-sentence basis, and for BLEU we set `effective_order=True` as recommended for sentence-level BLEU.

Typically, in real use with human translators, a TM may employ a match threshold, below which a sentence pair will not be returned. Our initial experiments using MT evaluation metrics forgo such a threshold, with the intention of comparing against MT without quality estimation. We also compare against the `FuzzyMatch-cli` implementation of Xu et al. (2020), which uses edit distance and suffix arrays to efficiently produce fuzzy matches from a TM.

The default setting of `FuzzyMatch-cli` does incorporate a threshold of a score of 0.8, along with minimal subsequence length of 3 and minimal subsequence ratio of 0.3. In many low-resource settings, these will not be met, resulting in no sentence pairs returned from the TM for a given source query. Thus we include in our initial experiments both a `FuzzyMatch`-default setting (as described above), and a `FuzzyMatch`-permissive setting, with a threshold of 0.0, a minimum subsequence length of 1, and a minimal subsequence ratio of 0.0. In practice, these parameters would likely result in useless sentence pairs being shown to the translator; here we use them to get a sense of just how low (according to automatic metrics) that quality may be. We also note that `FuzzyMatch-cli` is designed to be quite efficient; lowering these thresholds noticeably slows performance, which is another reason it would not be preferred in a real-life use case (or with larger datasets).

Due to the small size of most of our datasets, we forgo most optimizations in our BLEU and CHRF-based experiments. The only two optimizations that we employ are that we do deduplicate the translation pairs in the TM and, in the case of Inuktitut, we split the TM into smaller components, score them in parallel, and then recombine them.

Version	**BLEU**	**CHRF**
`FuzzyMatch`-default	0	6.3
`FuzzyMatch`-permissive	14.3	28.8
BLEU	**16.3**	31.9
CHRF	12.2	**38.2**
Oracle BLEU	22.3	36.1
Oracle CHRF	13.8	44.0

Table 3: Effect of different approaches to TM fuzzy matching, as measured by automatic metrics. The TM used is the DE-HSB training data, and results are scored against the DE-HSB test (devel_test) data. The final two lines show oracle performance.

3.2. TM Metrics and Evaluation

In their work, Simard and Fujita (2012) found a clear link between the MT metric used for TM similarity and the one used on the returned target language data for evaluation. That is to say, if one plans to evaluate with BLEU, it is typically best to use BLEU as the similarity metric for the TM extraction (or CHRF and CHRF, etc.). Our results replicate those findings.

The top portion of Table 3 shows the example of treating the `de-hsb` training data as the TM and extracting matches for the test data. The `FuzzyMatch`-default scores are so low because the vast majority of the sentences do not meet the threshold of 0.8; no match is returned for 1895 out of 2000 test sentences, whereas only one sentence does not have a match returned under the permissive setting. Consistent with Simard and Fujita (2012), the highest BLEU score occurs when using BLEU as the similarity function, while the highest CHRF score occurs when using CHRF as the similarity function. In our initial experiments across language pairs, this was typical; either the metric-matched version performed best, or they were tied. The one exception noted was from Inuktitut into English, where using CHRF as the similarity metric resulted in the highest BLEU score. This may be due to the morphological complexity of Inuktitut, which results in particularly low n-gram matches (but potentially high numbers of long sequences of matching characters).

3.3. Oracle

So far, we have compared source sentences s from the test set to source sentences s' from the TM, using their translations as the hypothesis t' (to be compared against the reference t). Of course, there is no guarantee that t' is in fact the sentence in the target side of the TM that is the closest to the reference t. For example, even if s and s' are identical, we could imagine a scenario where each word in t' is a paraphrase of a word in t, resulting in no overlap, such that there might be some other sentence in the TM that is closer to t. In real-life TM use, t is unknown (it is what is being produced by the translator). However, in these experiments, we can produce an *oracle*, i.e., the sentence t' that is closest to t according to the metric being used. This gives us a bit more insight into the performance of the TMs. As we see in Table 3, the oracles all outperform their realistic counterparts, to varying degrees (e.g., the oracle score using BLEU outperforms the realistic approach with BLEU by 6 BLEU points). It is worth noting, however, that the oracle experiments reinforce the finding that metric-matched scores outperform metric-mismatched scores: scoring the oracle BLEU output with CHRF results in a score (36.1) lower than that of scoring the *non-oracle* CHRF output with CHRF (38.2), and vice versa. In the following sections we will examine how the oracle and standard TM scores can help us examine mismatches between training, development, and test data.

4. Machine Translation Systems

We use existing machine translation systems to compare against the TM baselines. We make a distinction here between systems built using only the training data for the given language pair, those that incorporate the development data as well, and those that incorporate additional data.

For translation between Upper Sorbian and German, we use the "Bitext Baseline" systems described in Knowles and Larkin (2021). These are Transformer models (Vaswani et al., 2017) built using Sockeye (Hieber et al., 2018) with shared subword vocabularies of 10k and 15k subwords.

For the AmericasNLP language pairs, we use systems from Knowles et al. (2021), also Transformer models built using Sockeye. We look at that paper's baseline models (built using only the training data for a given language pair). In Section 6 we also examine multilingual finetuned models trained on the full training data (called S.1), and ensembled multilingual finetuned systems that also trained on the development data (called S.0). The multilingual systems all incorporate data from the four language pairs, but no other external data beyond the training and development data provided by the task organizers.

We also compare against state of the art (SOTA) systems, all of which incorporated additional data beyond that which was available to the TMs. The SOTA system is defined as the one that performed highest on the stated metric for the task (CHRF for AmericasNLP, a combination of metrics for Upper Sorbian and German, and human rankings for Inuktitut and English, with a caveat that the English to Inuktitut rankings are incomplete and were run only on Hansard data). The AmericasNLP SOTA systems are all from Vázquez et al. (2021), the Upper Sorbian systems are from Knowles and Larkin (2021), the Inuktitut to English systems are from Zhang et al. (2020) and the English to Inuktitut systems were from Hernandez and Nguyen (2020).

5. Analyzing Train/Dev/Test Mismatch

When training data, development data, and test data are all sampled from the same distribution, they would be considered well-matched. In that situation, we would reasonably expect that treating the (larger) training data as a TM would typically result in higher scores than treating the (smaller) development set as a TM, simply because there is a larger set of data to match against.

We can examine this with both oracles and with the realistic, non-oracle, setting. In Table 4, the TM rows show CHRF scores for realistic TM baselines built from development data (dev), training data (train), or their combination (all). For the well-matched case of HSB and DE, we see a large gap between the TM (dev) and TM (train) scores, with the larger TMs resulting in higher CHRF scores. The TM (all) score just slightly outperforms the TM (train). We also find this to be the case in the oracle setting, where oracle CHRF scores drop from 44.0 to 28.8 and from 46.2 to 32.7, respectively, when switching from the full train TM to the development set as TM.

There is a similar pattern for English and Inuktitut Hansard data, with the TM (train) outperforming the TM (dev), and no difference between the TM (all) and the TM (train). However, when looking at news data for this language pair, the TM (dev) scores actually outperform the TM (train) scores, indicating a closer match between the development data and test data than between the training data and the test data. This reflects the difference in domain – even though there is much more data available in the training set, it is not as well matched or may be missing certain vocabulary items, expressions, or structures that appear in the domain-matched development data.

In the AmericasNLP shared task, these differences are sometimes even more extreme. For the two languages with the smallest datasets, Rarámuri and Wixárika, the development data TMs outperform the training data TMs by 8.7 and 6.2 CHRF, respectively. We observe a similar pattern in the oracle setting (not shown in the table). We also note that the (non-oracle) development set TMs outperform the combined train and development set TMs for all four of these language pairs, with larger differences observed in the lower-resource pairs. Importantly, this indicates that for at least some of the sentence pairs in the test data, there exists a source sentence in the training data TM which is closer to the test sentence than any source sentence in the development data, but for which the target side is a *worse* match.

This reflects the range of challenges in low resource machine translation, which are not perfectly correlated: data size, domain differences, and dialect differences. While we would expect domain differences to be consistent on both the source and target side of a TM, dialect, orthography, or tokenization differences could affect just one language.

Through this simple lens, we can see that even though all of these are "low-resource machine translation", there is quite a bit of variation about what that means. Table 4 shows TM and MT (including both baseline and state-of-the-art) CHRF scores. Notably, when training Transformer models on just the training data used to build the TM (train), we see that in the well-matched scenarios (Upper Sorbian and German), the MT baseline improves dramatically over the TM, while in the mismatched scenarios (AmericasNLP) the baseline struggles to improve or even falls below some of the TM scores. In all cases, the state of the art systems incorporated additional monolingual or bilingual training data, so it is not immediately possible to tease apart model, architecture, and training improvements and data size increases (we discuss this in Section 7).

6. Fine Grained Analysis

Now we consider a fine-grained analysis of the translation of individual sentences. Using a TM (in the realistic, rather than oracle, setting), we can start to examine how well our machine translation systems are able to generalize. We typically evaluate our machine translation systems over a full test set, rather than looking at the performance of individual sentences.

6.1. Experiments

Here we'll consider each source sentence s in the test set, its corresponding hypothesis t' from the TM (in this section, we use the best-performing TM for the given language pair, namely the TM using training and development data for DE-HSB and the development data TM for ES-TAR), and the output t'' of a machine translation system. Given the reference t, we can compute the difference $\mathcal{M}(t'', t) - \mathcal{M}(t', t)$. A positive result indicates that the machine translation system outperformed the TM, a difference of zero means they performed equivalently, and a negative score indicates that the TM output outperformed the MT for the given sentence.

In our well-matched scenario (DE-HSB), simply training a baseline system results in the vast majority of differences being positive. Even for sentences that had very low match scores from the TM, we are able to produce improved translations. This means that the MT system has successfully generalized from the training data to novel sentences. Figure 1 shows this. Given a TM consisting of training data and development data, 93.9% of the time, the *baseline* MT output is as good as or better than the TM fuzzy match (as measured by CHRF); these are represented by all dots at or above 0 on the y-axis.

However, for the least well-matched scenario, Spanish-Rarámuri (ES-TAR), that is not the case. A baseline system trained on the training data only scores better than (or equal to) the TM 12.0% of the time, meaning that more than 8 times out of 10, you would be better off choosing the TM output than the MT output. Training a multilingual model and finetuning (without use of development data, the S.1 model) improves on this somewhat, with the MT system performing equivalently to or better than the TM 16.9% of the time. Incorporating the development data into the training (S.0 model) improves things; the new MT system outperforms or equals the TM on 62.7% of the sentences in the test set. This still means that in more than 1/3 cases, it would have been better to choose the output of the TM.

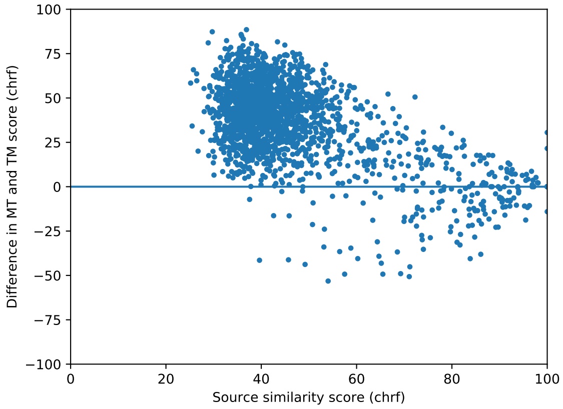

Figure 1: Difference between sentence-level CHRF scores for MT baseline and TM matches, by source side match score, German to Upper Sorbian. Greater x values indicate higher source side similarity according to CHRF, while greater y values indicate a greater improvement of the MT system over the TM output. (Negative y values indicate that the TM output scored higher than the MT. Note that for higher source match scores, there is typically less room for improvement.)

The CHRF scores of those systems are as follows: baseline 14.0, S.1 model 14.3, and S.0 model 24.7. We note that this S.0 model (the best for which we had the MT output available) is the third-best system submitted to the AmericasNLP task for this language pair, outperformed by 1.1 CHRF by the top-performing submission (Vázquez et al., 2021), which incorporated additional external parallel Spanish-Rarámuri data. Figure 2 shows both the baseline and S.0 models, with the tail of each arrow starting at the baseline point and the arrow head at the S.0 model value, showing the trend of improvement at the sentence level. There is no clear correlation between source metric scores and whether it would be better to use TM or MT output, even for the Spanish-Rarámuri system trained solely on the training data.

6.2. Discussion

Of course, the real proof of this would be to perform human evaluations comparing the usefulness of MT and TM output. There are various reasons to expect that such an experiment may not be perfectly correlated with these automatic results. First, scoring individual sentences (as opposed to full test sets) is known to be

System	es-hch	es-tar	es-nah	es-gn	hsb-de	de-hsb	iu-en (H)	en-iu (H)	iu-en (N)	en-iu (N)
TM (dev)	25.2	21.0	23.2	20.2	27.7	23.9	28.7/21.3	19.6/15.0	25.0/22.8	19.1/14.6
TM (train)	19.0	12.3	20.0	17.3	40.6	38.2	32.7	21.8	24.8	15.8
TM (all)	24.4	18.3	21.4	19.0	40.7	38.3	32.7	21.8	26.0	16.9
Base. MT	20.5	14.0	19.1	22.3	74.7	74.8	-	-	-	-
SOTA MT	36.0	25.8	30.1	37.6	79.8	79.6	50.9	35.0	25.4	30.0

Table 4: TM systems (using CHRF for source similarity over the training data or the training data and the development data; non-oracle) scored using CHRF, along with baseline and state of the art (SOTA) MT output. Note that hsb-de and de-hsb scores are for the devel_test set, rather than the test set (the SOTA MT system is the one that performed best on test at the shared task, used here to decode devel_test).

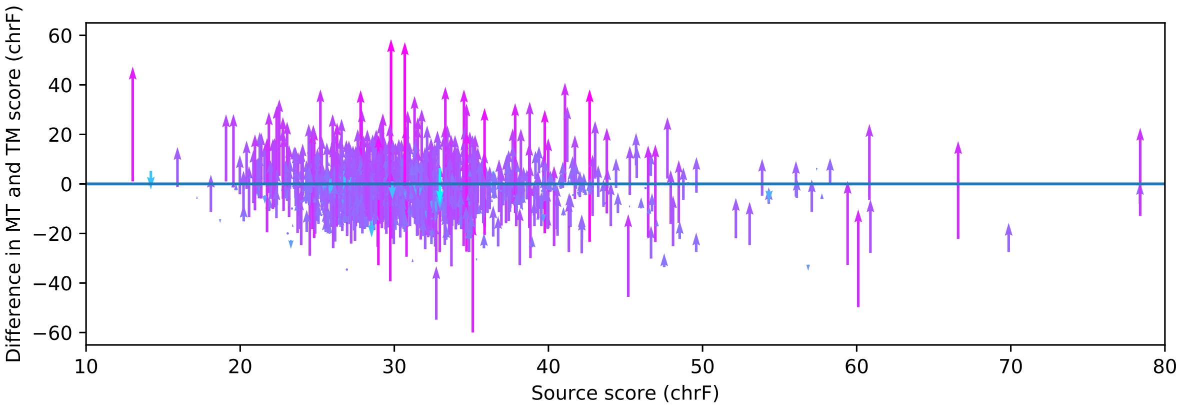

Figure 2: Improvement in difference between MT output and TM match of S.0 MT system over baseline at the sentence level, by source side match score (CHRF) for Spanish-Rarámuri. Pink (arrows pointing up) indicate positive improvement, cyan (arrows pointing down) indicate lowered performance.

quite difficult and noisy. Second, individual translators vary greatly in what tools they find helpful, both in the sense that translators often want computer aided translation interfaces that they can customize to their own preferences (Moorkens and O'Brien, 2017; Cadwell et al., 2018) and in the sense that inter-translator differences in the usefulness of machine translation (e.g., for post-editing) are often greater than the differences between several machine translation systems (Koehn and Germann, 2014). Finally, there are likely to be qualitative differences between the MT and TM output – we expect the TM to contain fluent sentences only, for example, while low resource MT may not be fluent (see Mager et al. (2021) for example).

Other ways that researchers can make use of TM outputs is through various analysis and visualization tools, like Compare-MT (Neubig et al., 2019) and MT-ComparEval (Klejch et al., 2015). In scenarios where researchers speak only the source language, looking at the source side TM matches can also provide, to some extent, a qualitative upper bound on the possible quality of output. For example, if the TM quality and MT quality perform poorly, by automatic metrics, a researcher could examine the source side TM matches. This could provide insight as to whether the matches are semantically related, or simply happen to have coincidental matching strings (as seen in the appendix of Knowles et al. (2021)).

There are limitations to this TM-based approach. It is computationally expensive as compared to simpler monolingual comparisons like n-gram overlap or language model perplexities (though not necessarily more so than building MT baselines). In this work, we have focused on language pairs with relatively small datasets, which renders this more manageable; for large datasets, one might wish to use a more highly-optimized tool like `FuzzyMatch-cli`. There are also no guarantees that it will capture all potential dataset problems.

We argue that viewing translation improvements over TM output – while an insufficient replacement for human evaluation – provides an additional perspective that can help researchers understand the gains that their MT systems are making (or the lack thereof).

7. Recommendations

We suggest that shared task organizers may wish to produce TM baselines, for several reasons. Compared to most machine translation systems, TM baselines that use an evaluation metric as a similarity measure are (nearly) non-parametric. If a shared task plans to use a particular metric for evaluation, we suggest using that exact metric (in the configuration that will be used for evaluation, along with any tokenization/detokenization that will be used in evaluation) for building a TM baseline.[4]

This is certainly not to discourage the use of additional machine translation baselines, but it could help control for a number of issues in machine translation baselines. In particular, it avoids the issue of "lucky" or "unlucky" runs and initializations.[5] Additionally, it has less of a risk of variance based on the baseline builder's experience (e.g., someone with much expertise in parameter tuning building an impossible-to-beat baseline vs. someone with less experience building a baseline that is "too easy" to beat).

Over time, TM baselines with a fixed metric used for similarity scores would also help to elucidate differences between algorithm-based and data-based improvements to MT quality. For example, if a task is repeated several years in a row, are improvements in baseline scores due primarily to increases in data, improvements to machine translation algorithms, or some combination of the two? Similarly, in shared tasks where additional data collection is allowed, participants could also compute TM baselines over their data collections, providing additional insight into the value of the data resources they have built. These could then be more easily compared across participants, without the confounds of different training and modeling decisions.

8. Conclusions

We propose translation memory baselines as a complement to machine translation baselines and human analysis. Human evaluations are, of course, a more meaningful way to measure improvements in machine translation, particularly for low resource languages. However, it can be challenging to perform human evaluations, and if translation quality is extremely low, it may be quite difficult to accurately judge MT adequacy (e.g., output may be too disfluent for the concept of adequacy to be applied). Comparing against a translation memory baseline could have several benefits: it can provide insights about mismatches between datasets, it is more appropriate than comparing against a baseline of "nothing", and it can provide rough estimates of how an MT system might perform against a translation memory in a CAT setting.

Acknowledgements

We thank the anonymous reviewers for their helpful comments and suggestions. We thank our colleagues Chi-kiu Lo and Michel Simard for discussion, Gabriel Bernier-Colborne for his comments on the paper, and Darlene Stewart and Samuel Larkin for their contributions to the previously-published MT systems used in these experiments.

9. Bibliographical References

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

Cadwell, P., O'Brien, S., and Teixeira, C. S. C. (2018). Resistance and accommodation: factors for the (non-) adoption of machine translation among professional translators. *Perspectives*, 26(3):301–321.

∀, Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabongo Kabenamualu, S., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Espoir, M., van Biljon, E., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B., Bassey, B., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, November. Association for Computational Linguistics.

Fraser, A. (2020). Findings of the WMT 2020 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 765–771, Online, November. Association for Computational Linguistics.

Hernandez, F. and Nguyen, V. (2020). The ubiqus English-Inuktitut system for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 213–217, Online, November. 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 Track)*, pages 200–207, Boston, MA, March. Association for Machine Translation in the Americas.

[4] If translation is only being performed in one direction, and there are reasons to differ in metric parameters across languages, one could either choose to use an oracle TM or choose reasonable parameters for the source side. Either way, this is likely to have fewer parameters to manage and report than a machine translation baseline.

[5] Though the standard response to this is to train multiple baselines, this may not always be possible.

Joanis, E., Knowles, R., Kuhn, R., Larkin, S., Littell, P., Lo, C.-k., Stewart, D., and Micher, J. (2020). The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2562–2572, Marseille, France, May. European Language Resources Association.

Klejch, O., Avramidis, E., Burchardt, A., and Popel, M. (2015). MT-compareval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, (104):63–74.

Knowles, R. and Larkin, S. (2021). NRC-CNRC systems for Upper Sorbian-German and Lower Sorbian-German machine translation 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 999–1008, Online, November. Association for Computational Linguistics.

Knowles, R., Stewart, D., Larkin, S., and Littell, P. (2021). NRC-CNRC machine translation systems for the 2021 AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 224–233, Online, June. Association for Computational Linguistics.

Knowles, R. (2021). On the stability of system rankings at WMT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 464–477, Online, November. Association for Computational Linguistics.

Koehn, P. and Germann, U. (2014). The impact of machine translation quality on human post-editing. In *Proceedings of the EACL 2014 Workshop on Humans and Computer-assisted Translation*, pages 38–46, Gothenburg, Sweden, April. Association for Computational Linguistics.

Libovický, J. and Fraser, A. (2021). Findings of the WMT 2021 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 731–737, Online, November. Association for Computational Linguistics.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online, June. Association for Computational Linguistics.

Moorkens, J. and O'Brien, S. (2017). Assessing user interface needs of post-editors of machine translation. *Human issues in translation technology*, pages 109–130.

Neubig, G., Dou, Z., Hu, J., Michel, P., Pruthi, D., Wang, X., and Wieting, J. (2019). compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

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, July. 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, September. 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, Brussels, Belgium, October. Association for Computational Linguistics.

Simard, M. and Fujita, A. (2012). A poor man's translation memory using machine translation evaluation metrics. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas*. Association for Machine Translation in the Americas.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Isabelle Guyon, et al., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Vázquez, R., Scherrer, Y., Virpioja, S., and Tiedemann, J. (2021). The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online, June. Association for Computational Linguistics.

Xu, J., Crego, J., and Senellart, J. (2020). Boosting neural machine translation with similar translations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1580–1590, Online, July. Association for Computational Linguistics.

Zhang, Y., Wang, Z., Cao, R., Wei, B., Shan, W., Zhou, S., Reheman, A., Zhou, T., Zeng, X., Wang, L., Mu, Y., Zhang, J., Liu, X., Zhou, X., Li, Y., Li, B., Xiao, T., and Zhu, J. (2020). The NiuTrans machine translation systems for WMT20. In *Proceedings of the Fifth Conference on Machine Translation*, pages 338–345, Online, November. Association for Computational Linguistics.

10. Language Resource References

Joanis, Eric and Knowles, Rebecca and Kuhn, Roland and Larkin, Samuel and Littell, Patrick and Lo, Chi-kiu and Stewart, Darlene and Micher, Jeffrey. (2020). *The Nunavut Hansard Inuktitut–English Parallel Corpus 3.0 with Preliminary Machine Translation Results*. European Language Resources Association.

Libovický, Jindřich and Fraser, Alexander. (2021). *Findings of the WMT 2021 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT*. Association for Computational Linguistics.

Mager, Manuel and Oncevay, Arturo and Ebrahimi, Abteen and Ortega, John and Rios, Annette and Fan, Angela and Gutierrez-Vasques, Ximena and Chiruzzo, Luis and Giménez-Lugo, Gustavo and Ramos, Ricardo and Meza Ruiz, Ivan Vladimir and Coto-Solano, Rolando and Palmer, Alexis and Mager-Hois, Elisabeth and Chaudhary, Vishrav and Neubig, Graham and Vu, Ngoc Thang and Kann, Katharina. (2021). *Findings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.