The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance

Cyril Goutte **Marine Carpuat** **George Foster**
Interactive Language Technology
National Research Council Canada
Gatineau, QC, J8X 3X7
Cyril.Goutte@nrc.ca Marine.Carpuat@nrc.ca George.Foster@nrc.ca

Abstract

When parallel or comparable corpora are harvested from the web, there is typically a trade-off between the size and quality of the data. In order to improve quality, corpus collection efforts often attempt to fix or remove misaligned sentence pairs. But, at the same time, Statistical Machine Translation (SMT) systems are widely assumed to be relatively robust to sentence alignment errors. However, there is little empirical evidence to support and characterize this robustness. This contribution investigates the impact of *sentence* alignment errors on a typical phrase-based SMT system. We confirm that SMT systems are highly tolerant to noise, and that performance only degrades seriously at very high noise levels. Our findings suggest that when collecting larger, noisy parallel data for training phrase-based SMT, cleaning up by trying to detect and remove incorrect alignments can actually degrade performance. Although fixing errors, when applicable, is a preferable strategy to removal, its benefits only become apparent for fairly high misalignment rates. We provide several explanations to support these findings.

1 Introduction

Parallel or comparable corpora are routinely harvested from the web or other large resources using automatic or semi-automatic methods (Tiedemann, 2009; Callison-Burch et al., 2009; Fung et al., 2010). Although this enables rapid collection of large corpora, it also requires fairly automated systems, which process large amounts of data with little human supervision. In particular, corpora are segmented into sentences which are then aligned using automatic sentence alignment techniques, e.g. (Moore, 2002). Despite the good performance of state-of-the art automatic sentence alignment, the size of the corpora and the cascading of earlier document or paragraph matching steps tend to result in many misaligned sentence pairs. We estimate in section 3.1 that three of the corpora used in the WMT evaluation[1] contain between 1.2% and 13.1% misaligned sentence pairs.

It is often argued that SMT systems, due to their statistical nature, are relatively robust to sentence alignment errors. However, there is to our knowledge little empirical support for this belief (Khadivi and Ney, 2005). In this paper, we attempt to address this by analysing the impact of sentence misalignment rate on SMT output quality. We provide three main contributions:

1. We describe a sampling approach for estimating alignment noise reliably with limited human effort, and provide resulting estimates on common corpora;

2. We measure the robustness of a typical phrase-based MT system to increasing levels of misalignment errors by simulating these errors in a high-quality corpus;

3. We suggest a strategy for training MT systems from noisy data coming for example from parallel or comparable corpora, relying on the two previous contributions.

[1] http://statmt.org/wmt12/translation-task.html

We emphasize that this work deals with *sentence alignment errors* as opposed to word alignment errors, although we later discuss the interaction between them.

Our investigations require a large, high quality, sentence aligned corpus, into which we add increasing amounts of random misalignments (sections 2-3). Section 4 presents the experimental results obtained by training a state-of-the-art phrase-based SMT (PBMT) system on the degraded corpora and measuring the impact on its performance. We show that indeed, PBMT systems are surprisingly robust to alignment noise, and that performance is actually higher on the noisy corpus than on a clean version of the corpus where alignment errors have been filtered out. Section 5 discusses these findings.

2 Data

In order to investigate the impact of alignment error on SMT performance, we need a parallel corpus that is *large enough* to be representative of the large data conditions needed to train state-of-the-art SMT systems, and *clean enough* to let us control the level of sentence-alignment noise. Unfortunately, as we demonstrate in the next section, most widely available large corpora have moderate-to-high misalignment rates, making it impossible to obtain a "clean" reference for the SMT performance. The Europarl corpus appears to have high sentence alignment quality, but it is fairly small by current standards, at least on well-studied language pairs. We acquired a large corpus of French-English parallel data from the Canadian Hansard, including proceedings from the House of Commons and from committees.[2] Using careful alignment, we obtained a total of 8,194,055 parallel sentences. We reserved subsets of 16,589 and 17,114 sentences in order to sample development and test sets, respectively, leaving up to 8,160,352 sentences for training the PBMT system. Section 3.1 shows that the estimated misalignment rate for this corpus is 0.5%, and section 3.2 describes how we gradually introduce increasing amounts of alignment error into the corpus for the purpose of our experiments.

[2]This corpus is available on request.

3 Method

We first introduce a sampling method for estimating the baseline level of alignment error in a parallel corpus, and apply it to the Hansard corpus, as well as several others for comparison. We then artificially introduce random alignment errors into the Hansard, as described below. We also briefly describe the PBMT system used in the experiments.

3.1 Estimating Sentence Alignment Error

We model the estimation of alignment errors using a simple binomial model. In this model, the bilingual corpus containing well-aligned and misaligned sentence pairs can be viewed as a large "urn" containing black and white balls. By sampling sentence pairs and evaluating whether they are correctly aligned or not, we draw balls from the urn and look at their colour. The outcome of this experiment is used to estimate the rate of misalignment in the corpus, just as we would estimate the proportion of white balls from our draw.

Let S be the number of sentence pairs that we sample, out of which "m" are misaligned, and $S-m$ correctly aligned. Given the (unknown) misalignment rate μ, the distribution of the number m of misaligned pairs is given by a binomial

$$P(m|\mu, S) = \frac{S!}{m!(S-m)!}\mu^m(1-\mu)^{S-m}.$$

A natural choice for the prior distribution on μ is a symmetric Beta, $p(\mu) = \frac{\Gamma(2\lambda)}{\Gamma(\lambda)^2}\mu^{\lambda-1}(1-\mu)^{\lambda-1}$. With $\lambda = 1$ this is a uniform distribution, while $\lambda = 1/2$ is the non-informative prior (Box and Tiao, 1973), which we will use here. Bayes' formula yields the posterior distribution for μ:

$$p(\mu|S, m) \propto \mu^{m-\frac{1}{2}}(1-\mu)^{S-m-\frac{1}{2}},$$

which is again a Beta distribution, with parameters $m + \frac{1}{2}$ and $S - m + \frac{1}{2}$.

From the posterior, we can derive a number of interesting estimates such as the expected misalignment rate, $\widehat{\mu} = (m + \frac{1}{2})/(S + 1)$, which is a smoothed version of the empirical estimate. More importantly, we can use the posterior distribution to derive confidence intervals and guarantees on the maximum misalignment rate.

Corpus	S	m	$\widehat{\mu}$ (%)	95% CI
Europarl	300	3	1.2	[0; 2.3]
UN	300	8	2.8	[0; 4.5]
Giga	300	39	13.1	[0; 16.4]
Hansard	300	1	0.5	[0; 1.3]

Table 1: Estimated expected misalignment rate ($\widehat{\mu}$) for four MT corpora. S is the number of sentence pairs evaluated, and m the number of incorrectly aligned ones. The confidence interval is the one-sided 95% interval.

In order to illustrate this, we consider three corpora from among the official WMT dataset, in addition to the Hansard described above:

- Europarl: 1,328,360 sentence pairs;
- United Nations: 12,886,831 sentence pairs;
- Giga (10^9) corpus: 22,520,400 sentence pairs;
- Hansard: 8,160,352 sentence pairs.

We sample a small number of sentence pairs (usually 300) and manually evaluate the correctness of the alignments. The results are given in Table 1. Depending on the corpus, between 1 and 39 pairs were found to be misaligned, resulting in expected misalignment rates between 0.5% and 13%.

The differences between these corpora is illustrated on Figure 1 where we plot the posterior distribution resulting from our evaluation. We see that the estimated misalignment rates as well as the uncertainty on this estimate (the spread of the posterior distribution) vary widely. Note that the expected misalignment rate does not coincide with the location of the highest probability (mode) of the distribution, which is normal for a skewed distribution.

Giga is the largest corpus, but also has the highest misalignment rate, at an estimated 13%, which corresponds to close to 3 million incorrect sentence pairs in this corpus (more than twice the entire Europarl corpus). The UN corpus has a lower misalignment rate, estimated below 3%. Europarl is even cleaner, with $\widehat{\mu} = 1.2\%$, but also much smaller. Finally the Hansard has an estimated misalignment rate of around 0.5%. The last column shows that, for the Hansard corpus, we can say with 95% confidence that the misalignment rate is below 1.3%. The estimated 0.5% misalignment corresponds to around

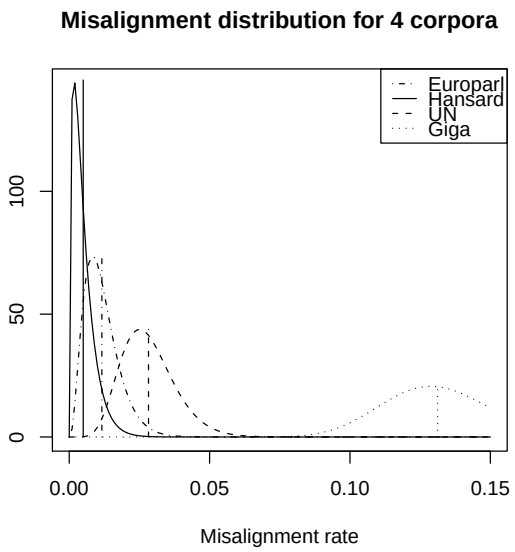

Figure 1: Estimated distribution of misalignment rates.

40k incorrectly aligned sentence pairs out of this 8.16M sentence corpus. This motivates our use of that corpus as a basis for our experiments: it is relatively large and very "clean".

The overall message of this section is that using a sampling-based approach, it is possible to obtain a fairly reliable bounded estimate of the misalignment quality of a corpus with relatively modest effort.

3.2 Introducing Alignment Errors

Starting from the large, high-quality Hansard corpus, we gradually introduce random alignment errors, by increments of 10%. We randomly sample a number of sentence pairs corresponding to the target error level,[3] and do a permutation of the target (English) side. For example, for 10% noise, we sample around 775,000 sentence pairs, then the target side of the first pair in the sample is assigned to the source side of the second pair, the target of the second to the third source, etc. We also ensure that each perturbation is strictly included in a larger perturbation, ie the 20%-noise misalignments contain the 10%-noise misalignments, etc. To average results over the choice of alignment errors, we sample 6 random samples at each of 10%, 20%, ... 90% misalignment rate, hence a total of 54 noisy corpora, plus the original, "clean" one.

As we introduce noise by sentence permuta-

[3] Minus 0.5% to account for the baseline misalignment rate.

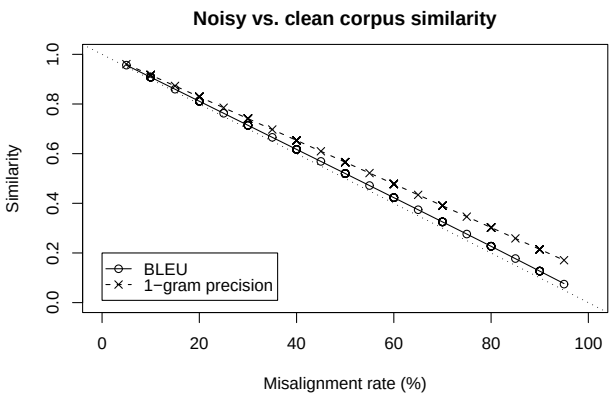

Figure 2: Similarity (BLEU and 1-gram precision) of noisy corpora vs. clean reference.

tion within a well-defined domain, some words or phrases may still align to source words. We quantify this by computing the overlap between the noisy and original versions of the target side at each noise level, using BLEU and 1-gram precision. Figure 2 shows that the overlap between noisy and clean corpora decreases linearly and roughly matches the percentage of clean sentences in the corpora. This suggests that there are only few matching words between permuted and original target sentences, hence little chance of extracting correct phrase pairs from incorrect alignments. This is discussed further in section 5.

As additional alignment errors are introduced artificially, we know exactly which pairs are misaligned, apart from the 0.5% baseline errors. To support further experiments, we produce "cleaned" versions of the corpora at each noise level, where we remove the artificially introduced errors, leaving only the unmodified sentence pairs. These have the same baseline misalignment rate of 0.5% but are smaller in size (90.5%, 80.5% . . . 10.5% of the full corpus). Although we can't remove the final 0.5% of misaligned sentence pairs, for convenience we call this "perfect filtering" below.

3.3 Training the SMT model

For each sample at each perturbation level, we train a standard PBMT model and estimate its performance on the reference test corpus. We use a typical PBMT system which has achieved competitive results in recent NIST and WMT evaluations (Larkin et al., 2010). We use the following feature functions in the log-linear model:

- 4-gram language model with Kneser-Ney smoothing (1 feature);
- relative-frequency and lexical translation model probabilities in both directions (4 features);
- lexicalized distortion (6 features); and
- word count (1 feature).

The parameters of the log-linear model are tuned by optimizing BLEU on the development set using MIRA (Chiang et al., 2008).[4] Phrase extraction is done by aligning the corpus at the word level using both HMM and IBM2 models, using the union of phrases extracted from these separate alignments for the phrase table, with a maximum phrase length of 7 tokens. Phrase pairs were filtered so that the top 30 translations for each source phrase were retained. The translation performance was measured using BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005).

4 Results

We show how SMT output degrades with increasing alignment noise. We see that, surprisingly, even relatively high levels of noise have little impact on translation performance. We then compare the robustness of PBMT systems to that of Translation Memories, a commom computer-aided translation tool.

4.1 Impact on translation performance

Figure 3 shows how translation performance, as estimated by BLEU (circles), degrades when the number of misaligned sentence pairs increases. Not surprisingly, increasing the noise level produces a general decrease in performance. Although there are variations depending on the samples, the smoothed curve (solid line) is strictly decreasing as expected. What may be more surprising is how little the performance is affected as alignment error approaches relatively high levels. After adding 30% alignment errors, the average BLEU score drops from the 37.59 obtained on the "clean" corpus, down to 37.31, less

[4] MERT gives qualitatively similar results.

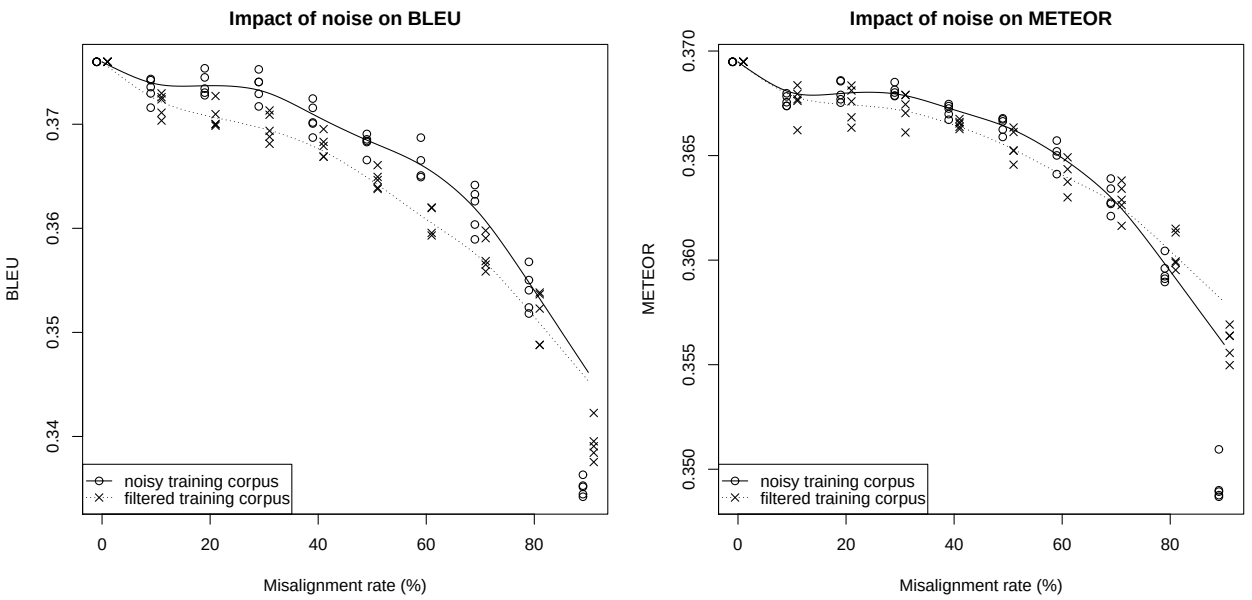

Figure 3: Impact of sentence misalignment error on SMT performance estimated with BLEU (left) and METEOR (right). Results are for training on full corpora (o) and when misalignment errors are filtered out (x). Curves are smoothed using a local linear smoother (`locpoly` in R).

than 0.3 BLEU points below. Translation performance degrades faster after that, but only takes a large hit when the misalignment rate grows beyond 60-70%, ie far more incorrect alignments than correct ones, a situation that should be very easy to detect (and hopefully rare) in practice. Note that the average BLEU score at 70% noise is still 36.13, less than 1.5 BLEU points below the "clean" BLEU.

In order to show that these results are not an artefact of using BLEU for estimating the translation quality, we also produce curves showing the impact of misalignment noise on METEOR (Banerjee and Lavie, 2005), another popular metric. The right plot in Figure 3 shows these results. We see that although the metrics are different, the general picture is quite similar: low to moderate noise has little to no impact on the performance estimated by METEOR, apart from the initial variability due to the fact that we have only one training set at 0.5% error. Performance starts to really degrade from around 30% noise, and gets much worse after 60-70% noise.

4.2 Comparison with perfect filtering

To put this into perspective, we perform another set of experiments on training corpora where we filter out the misaligned pairs that we introduced earlier. This results in high quality but smaller corpora of 90%–10% of the original corpus size. The performance of the PBMT systems trained on these "filtered" corpora is plotted as crosses and dotted line in Figure 3. The surprising outcome of this experiment is that the performance on the filtered corpus is no better than when misaligned sentences are kept in the training data. In fact, this "perfect filtering" produces a small but consistent *decrease* in performance until very high noise levels are reached. One explanation for this is that the increase in quality in the *translation model* that we expect to result from the filtering is insufficient to compensate for the decrease caused by the reduced amount of data available for training the *language model*.

In order to test that hypothesis, we trained a hybrid model at each noise level, using the entire corpus for training the language model, and the filtered (cleaner but smaller) corpus for training the translation and distortion models. The language model uses only the target side of the corpus, and is invariant through the permutations used to introduce the noise in our experiments. The results plotted in Figure 4 completely validate our explanation: the use of a language model trained on a larger corpus greatly improves over the "filtered" performance: The hybrid (dashed) curve is always above the "filtered" (dotted) curve. Note also that until around 30%

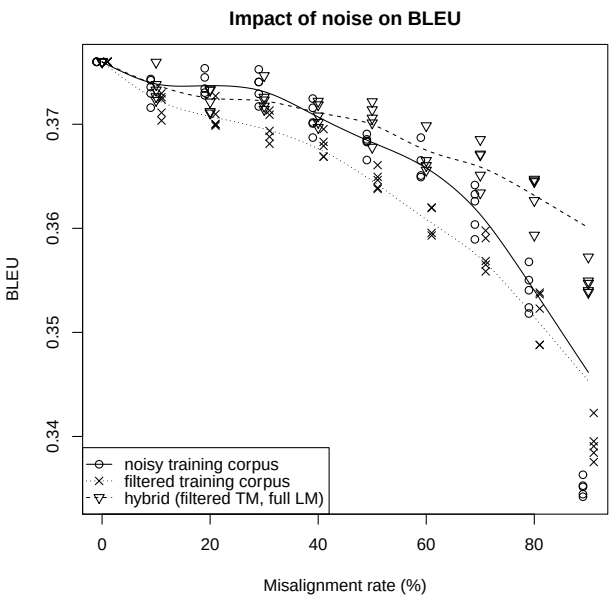

Figure 4: BLEU score vs. misalignment error, for PBMT systems trained on full corpora (o), filtered corpora (x) and hybrid approach (triangles). For clarity, we introduced a slight horizontal offset for markers.

noise, performance of the "noisy" (solid) and "hybrid" (dashed) models is very close, showing that the translation model does not suffer much from being trained on a corpus containing up to 30% sentence alignment errors. Above 40% noise, the hybrid model dominates the other two quite clearly.

The "perfect filtering", removing material from both TM and LM training, may seem like an odd choice given the previous analysis. This is however the typical way large corpora are built and made available. For example, among the large corpora used for the WMT workshop shared tasks, most of them were harvested (semi-)automatically; none includes monolingual versions that incorporate filtered out material.

4.3 Comparison with Translation Memory

We contrast the impact of noise on PBMT with its impact on the output of a simple translation memory (TM). Translation memories are a Computer-Aided Translation tool widely used by translators. Although they are only used in practice to provide translations to test segments that are highly similar to the content of a bilingual corpus, they may be generalized to provide what is essentially a one-nearest-neighbour translation prediction. Our TM implementation searches the parallel training corpus for the source sentence with maximum similarity[5] to each test sentence, and outputs the corresponding target sentence. TM quality is much worse than SMT: the TM applied to the clean corpus yields a BLEU score of 13.64. It is also much more sensitive to sentence alignment errors: at 30% and 90% noise levels, the average scores for the TM are respectively 9.75 and 1.97 (relative decrease of 28% and 86%). This confirms the remarkable robustness of SMT, for which BLEU scores degrade much more gracefully.

5 Discussion

Although it is well known that PBMT systems are robust to noise, our results indicate that this robustness holds to a remarkable extent, in fact to levels of noise that are far higher than usually found in parallel corpora. Three questions are relevant: what is the explanation for this phenomenon; can we expect these results to generalize to other settings and to non-synthetic noise; and what are the practical ramifications for training PBMT systems?

5.1 Analysis of SMT Robustness

To explain our results, we consider the effect of alignment noise on the phrase table. For a typical entry, $P(t|s)$ is peaked on a few values of t. Other incorrect target phrases appear erroneously in the phrase table due to sentence and word alignment errors. As we introduce more alignment errors, the number of target segments incorrectly associated with a source segment grows, and the estimated probability of the correct translations drops correspondingly. However, due to the random nature of the alignment errors, incorrect translations keep a low probability: $P(t|s)$ gets "broader" and "flatter", but as long as there are enough good alignments to collect statistics, the most likely t are still correct and translation is only moderately affected. Note also that the phrase extraction heuristics may also help reject incorrect translations as they will only consider word alignments that satisfy a number of regularity constraints.

We support our explanation by inspecting the part of the phrase table that is used for producing trans-

[5] We use a smoothed BLEU score as the similarity metric.

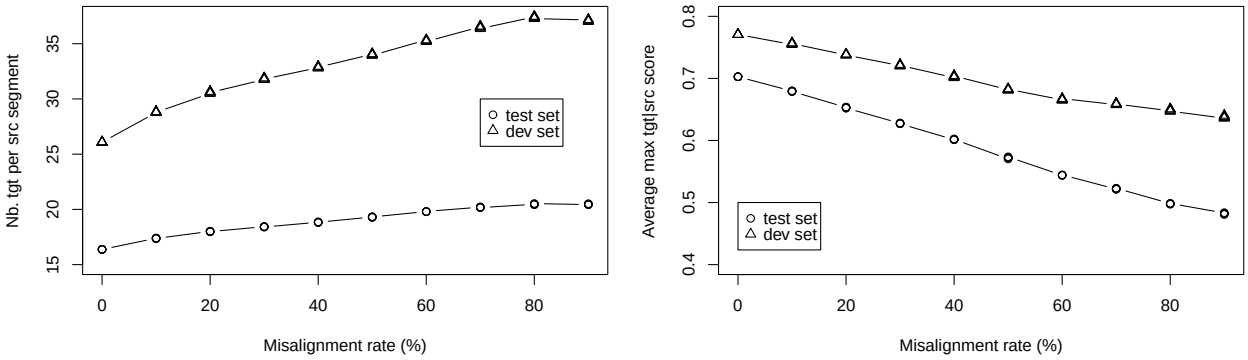

Figure 5: Average # target segment (left) and max. posterior probability (right) over source segments in phrase table.

lations on our test and dev sets, and computing two quantities that characterize the posterior $P(t|s)$: average number of target segments per source segment, and average maximum value of the posterior probability. Figure 5 shows that, as the number of misaligned sentences increases, the posterior indeed gets "broader" (there are more translations t for each s) and "flatter" (the probability of the most probable translation gets lower).

From a somewhat different perspective, the presence of sentence alignment errors in the training corpus may be seen as the addition of noise to what is essentially the input of the MT model estimation procedure. In machine learning, it has long been known that the presence of noise in the input data has an effect similar to regularization (Bishop, 1995). Tuned to the right level, it may therefore potentially yield improvements in performance. Although it is impractical to tune the appropriate level of sentence misalignment in a bilingual corpus, this, together with the smoothing effect illustrated above, helps shed further light on the robustness of PBMT to alignment errors.

5.2 Simulated Noise vs. Real Noise

A concern with all experiments that rely on simulation is how relevant the simulated conditions are to actually observed, real conditions. We first note that the noise we consider in this paper affects translation quality only—it assumes well-formed source and target sentences. Clearly, many real sources of noise will not have this profile, but these are arguably easier to detect and address, since detection can be based on monolingual properties (which will necessarily affect translation quality). However, we make no claims here about SMT robustness to this kind of noise.

The kind of noise we model can typically arise through automatic alignment procedures for parallel or comparable corpora. Our model makes two assumptions that do not always characterize real data. First, it assumes that errors are always "maximally bad", as demonstrated in figure 2: it is very unlikely that any valid phrase pairs will be extracted from mis-aligned sentence pairs. A substantial proportion of errors in real parallel and comparable corpora will not be this bad, and will permit extraction of some valid pairs from properly-aligned sentence fragments. Clearly these errors will not damage translation performance as much as our simulated errors. Our results can thus be seen as estimates of *minimum* noise robustness (when errors are counted at the sentence level).

A second assumption we have made is that noise will be uniformly distributed. This is a key assumption, as our analysis in the previous section shows. Clearly, real errors will sometimes exhibit systematic tendencies that violate this assumption. However, if these are frequent, they will be easy to detect and correct; and if they are infrequent, they will be inconsequential. The middle ground—"malicious" errors capable of tricking an SMT system into producing bad translations—seems implausible. Certainly we found no traces of any such effect in our manual alignment evaluations.

5.3 Training with Noise

The results in this paper, and the foregoing discussion, suggest a possible strategy for dealing with alignment noise when training an SMT system. The first step, as usual, is to manually inspect and address any obvious sources of noise. Since real corpora vary enormously, this step is difficult to automate. Next, estimate alignment error using the procedure in section 3.1. If this is greater than approximately 30%, discard low-confidence pairs until it is below 30%. (We assume the existence of sentence-pair confidence scores as a typical by-product of the alignment process.) When discarding sentence pairs, retain the target sentences for language model training.

We emphasize that this procedure is of course highly tentative. It will need to be adjusted for many different factors, such as language pair, the importance of small performance differences to the application, the reliability of confidence scores, etc. However, for large noisy corpora, or for small noisy corpora used as part of a larger training set, it offers significant potential speed and convenience advantages over the alternative of re-training from scratch and measuring performance at different noise levels.

6 Conclusion

We analysed the impact of sentence alignment errors on SMT quality. We first described a method for quickly estimating alignment noise by sampling, and provide estimates on common corpora. Then, through simulation, we analyzed the robustness of phrase-based MT to alignment errors.

Our results showed that phrase-based MT is highly robust to alignment errors: performance is hardly affected when the misalignment rate is below 30%, and introducing 50% alignment error brings performance down less than 1 BLEU point. We suggest that this may be thanks to the phrase table extraction and estimation procedure. Our results are limited to one corpus, language pair and SMT system, but suggest that efforts spent on elaborate procedures for filtering out sentence alignment errors from automatically harvested corpora may bring little payoff, or even degrade performance. More specifically, the increase in quality achieved by filtering out incorrect alignments may not offset the decrease resulting from lower corpus size.

These findings can inform strategies for training MT systems with noisy data. For instance, we suggest handling corpora with low alignment quality by filtering bilingual pairs so that alignment error is below 30% while keeping all target side segments for training the language model. This seems especially promising for sentence pairs extracted from comparable corpora, which we will investigate in future work.

References

S. Banerjee and A. Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

C. M. Bishop. 1995. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116.

G. E. P. Box and G. C. Tiao. 1973. *Bayesian Inference in Statistical Analysis*. Wiley.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Fourth Workshop on Statistical Machine Translation*, pages 1–28.

D. Chiang, Y. Marton, and P. Resnik. 2008. Online large-margin training of syntactic and structural translation. In *EMNLP*, pages 224–233.

P. Fung, E. Prochasson, and S. Shi. 2010. Trillions of comparable documents. In *LREC Workshop on Building and Using Comparable Corpora*, May.

S. Khadivi and H. Ney. 2005. Automatic filtering of bilingual corpora for statistical machine translation. In *Natural Language Processing and Information Systems*, LNCS 3513, pages 263–274. Springer.

S. Larkin, B. Chen, G. Foster, U. Germann, E. Joanis, J. H. Johnson, and R. Kuhn. 2010. Lessons from NRC's portage system at WMT 2010. In *5th Workshop on Statistical Machine Translation*.

R. C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In S. D. Richardson, editor, *AMTA*, volume 2499 of *Lecture Notes in Computer Science*, pages 135–144. Springer.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th Annual Meeting of the ACL*.

J. Tiedemann. 2009. News from OPUS—a collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, pages 237–248.