Clean Data for Training Statistical MT: The Case of MT Contamination

Michel Simard michel.simard@nrc.ca
National Research Council Canada
Multilingual Text Processing
1200 Montreal Rd., Ottawa, Ontario K1A 0R6 Canada

Abstract

Users of Statistical Machine Translation (SMT) sometimes turn to the Web to obtain data to train their systems. One problem with this approach is the potential for "MT contamination": when large amounts of parallel data are collected automatically, there is a risk that a non-negligible portion consists of machine-translated text. Theoretically, using this kind of data to train SMT systems is likely to reinforce the errors committed by other systems, or even by an earlier versions of the same system. In this paper, we study the effect of MT-contaminated training data on SMT quality, by performing controlled simulations under a wide range of conditions. Our experiments highlight situations in which MT contamination can be harmful, and assess the potential of decontamination techniques.

1 Introduction

As Statistical Machine Translation (SMT) systems are becoming more widely used in industry, we see and hear a growing amount of advice on "best practices" for SMT, most notably regarding system training and data collection and preparation. One recurring theme is that of "clean" data: how training data for SMT systems (which often comes in the form of translation memories) should be exempt of any "dirt", and how SMT users should go about "cleaning" their data. The most visible proponents of this clean-data approach are probably organizations such as TAUS[1], Asia Online[2] and PangeaMT[3].

Multiple forms of "dirt" are said to affect SMT quality. Asia Online (2009) discuss a number of examples: formatting codes and other kinds of markup, such as translation memory metadata; untranslated segments (i.e. translation units in which either the target segment is empty, or a copy of the source segment); segments containing inconsistent or mangled character encoding. "Badly" translated segments also fall into this category, although it is not quite clear what it means for a segment to be "badly" translated. Some of the examples given by the authors of the Asia Online study suggest that the presence of translator insertions can be considered "bad", as when, for example, the translator leaves a source-language term in parentheses for clarity. Non-literal translations are also frowned upon (e.g.: *What is a project file?* → *Présentation des fichiers projet*), differing usage of punctuation in the source and target languages, etc. The notion of dirt can extend, in some cases, to out-of-domain data, or even in-domain data containing terminology that is inconsistent with the current task. One translator's

[1] https://www.taus.net

[2] http://www.asiaonline.net

[3] http://pangeamt.com

treasure is another translator's trash.[4]

The kinds of problems mentioned in Asia Online's study are typical of SMT training data extracted from translation memories. Increasingly, however, SMT users are turning to the Web to complement their training sets (Resnik and Smith, 2003). There, other types of dirt can be expected to surface. One example is misaligned text: pairs of segments or documents in training data which are not translations of one another (Jiang et al., 2010; Goutte et al., 2012). In this case, the problem does not come from the data itself, but rather from the early processing steps by which the data is collected and organized for SMT training. While in typical translation memory usage, document- and segment-level alignments are validated manually, in Web-collected data, these steps need to be performed automatically. Goutte et al. (2012) estimate that some of the Web-collected data used in WMT evaluations[5] contain as much as 13% misaligned pairs of segments. Their study nevertheless reveals that typical SMT systems are relatively immune to this sort of dirt, and that even when as much as 30% of the training data is misaligned, SMT quality remains essentially unaffected.

Another kind of dirt comes in the form of machine translated documents: in order to make their online content available in multiple languages, some organizations are known to use unedited MT output. For example, Microsoft is known to have used MT extensively for some of their online technical documentation (Aikawa et al., 2007). More recently, EBay has started using MT for localizing product descriptions (Wohlsen, 2014). When automatically harvesting large amounts of bilingual documents from the Web, there is a growing risk that some unknown proportion of these documents will actually be machine translated.

In general, the presence of machine-generated data in training data that is otherwise assumed to have been produced by humans is seen as a bad thing. Using this kind of data to build systems aimed at simulating human behavior is likely to reinforce the errors committed by other systems, or even earlier versions of the same system. This is what motivates efforts such as those of Google to "watermark" the output of their MT systems (Venugopal et al., 2011), thus making it possible for them to recognize the output of their own systems, and to exclude it from their training sets.

In practice, however, it is not clear that machine-generated translations are always harmful in MT training data. In fact, in some situations, MT has been used specifically to compensate for the lack of adequate bilingual training data. For example, Ueffing (2006) proposed a self-training procedure, in which machine-translated versions of the test data itself was used as an additional source of knowledge in a phrase-based SMT system. This approach, as well as a number of variants, have shown to improve MT quality in different contexts (Bertoldi and Federico, 2009; Schwenk and Senellart, 2009; Lambert et al., 2011). In a somewhat related manner, (Madnani, 2010) and (Dyer et al., 2011) address SMT overfitting issues by adding machine-generated reference translations to their SMT system's tuning sets.

In the following pages, we focus on the effects of "accidental" MT contamination of MT training data: our goal is to assess to what extent this sort of contamination is harmful in SMT training, and what can be expected from cleaning procedures aimed specifically at eliminating this type of dirt. Our approach is to simulate MT contamination in otherwise clean data, and to measure the effect empirically on SMT systems trained with this data, using standard MT quality metrics. We describe the details of our experimental approach, the experimental data and the MT systems used in Section 2. The main results of our experiments are presented in Section 3, while the potential of data decontamination procedures is assessed in Section 4. Finally, we discuss the general implications of our findings in Section 5.

[4] *Weed* is possibly a better metaphor than *dirt* for many of these issues with SMT training data: "a wild plant growing where it is not wanted and in competition with cultivated plants" (from the Oxford US English Dictionary).

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

2 Method

2.1 General Approach

The general objective of this study is to assess the impact of MT contamination in SMT training data. Our experimental approach is to inject machine-translated output into an otherwise "clean" training set, then measure the effect on the quality of SMT systems trained with this data. The question we are addressing is whether or not one should worry about MT contamination: How should one handle a given subset of the available training data, knowing that it may contain an unknown proportion of MT output. Therefore all our experiments compare the performance of "baseline" systems, trained exclusively with "clean" data, with that of "augmented" systems, produced by adding *auxiliary training data* contaminated with MT. We consider systems augmented with varying amounts of auxiliary data, and containing varying degrees of MT contamination.

Our experiments also consider varying baseline conditions, in terms of quantity of baseline training data, text specialization and language pair. Because augmenting datasets from Web documents is a strategy that is mostly used in small-data settings, it makes sense to focus on scenarios where the baseline systems are relatively "small", trained with as little as 5000 segments in some cases, and not more than 200K sentence pairs. We consider data-collection efforts that aim at increasing training data by at least 50%, and as much as 8 times the size of the baseline dataset.

In the scenarios we wish to simulate, auxiliary data is collected from the Web. However, we wish to avoid biases resulting from differences between the baseline and auxiliary data's domain or genre. Therefore, in each scenario, experiments are performed using a single, relatively homogeneous pool of bilingual data, for which we have human-quality translations, and process it in different ways so as to simulate the desired conditions. More specifically, we machine-translate this data so as to obtain two different translations for each source sentence in the training data: one human translation, and one machine translation. Different experimental conditions can then be created that correspond to different mixes of human and machine-generated translations.

The quality of the MT in the auxiliary training data should be an important factor in our study: intuitively, we expect "bad" training MT to have a more profound adverse effect on MT quality than "good" training MT. However, this is a factor that is particularly hard to control for. As mentioned earlier, the scenario we are trying to simulate is one where auxiliary data is harvested from the Web. Within this kind of data, MT may have been produced under widely varying conditions. One reasonable assumption is that most of it has been produced using a "generic" (non-specialized or non-domain-adapted) system of decent quality, such as Google Translate. The problem with using Google Translate for our experiments is that we have no control on its training data: if we are to perform experiments using existing bilingual resources (corpora routinely used in research), we have no way of verifying that some or all of this data is not already used by Google to train their system. There are at least two solutions to this problem. One is to select the experimental data in such a way that we can guarantee it has not been used by Google Translate. The second is to use an MT system other than Google Translate, for which we have full control on the training data. For the experiments described here, we opt for the latter strategy. We describe our experimental data and the MT systems used below.

2.2 Data

We describe here the three distinct data sets used in our experiments. More details can be found in Table 1.

Dataset	Language	Segments	Source Tokens
Europarl	English-French	2M	55.5M
EMEA	French- English	837K	12.7M
GALE	Chinese-English	547K	17.7M

Table 1: Experimental data

Europarl This is release v7 of the well-known Europarl corpus (Koehn, 2005), in English and French. We arbitrarily set the source language to be English. We set aside 2000 randomly picked pairs of segments for tuning purposes. The test set for this dataset is the Europarl test set from the WMT 2008 shared task (Callison-Burch et al., 2008).

EMEA This is release 0.3 of the EMEA dataset, from the OPUS Corpus (Tiedemann, 2009) in English and French. In this case, we chose French to be the source language. This is a parallel corpus made out of PDF documents from the European Medicines Agency. While highly technical, it is very repetitive by nature. The data is organized into individual documents. We respected these natural divisions when subsampling the data: we randomly picked 60 documents from the complete dataset, and assigned 30 for tuning and 30 for testing. We took at most 100 segments from each document, to avoid having a few documents overwhelm the tuning or test sets. This procedure yielded 2185 and 1832 tuning and test segments respectively.

GALE This is a collection of Chinese-English corpora from the DARPA GALE initiative, the larger part coming from the FBIS corpus. For tuning and test, we used the NIST 04 (1788 segments) and NIST 05 (1082 segments) Chinese-English test sets respectively. Four reference translations are provided for each of these sets.

2.3 MT Contamination

In our experimental setup, MT systems play two different roles: they are first used as “contaminators”, i.e. to produce machine-translated target-language versions of training segments; then, they are used as “machine learners” in the rest of the study.

To generate MT contamination in the GALE dataset, we used an older version of SYSTRAN’s Chinese-English machine translation system (Yang et al., 2003). This is a rule-based system, customized for the domains of science and technology. It uses a rule-based word-segmenter and a bilingual lexicon with about 1.2M entries, containing words, expressions and rules.

For the English-French datasets (Europarl and EMEA), MT contamination was generated using SMT systems based on *Portage* (Larkin et al., 2010), the NRC’s phrase-based SMT technology. English-French and French-English systems were trained using a very large corpus of Canadian government data harvested from the Web (domain `gc.ca`), containing over 500M words in each language. Phrase extraction was 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. The following feature functions are used in the log-linear model: 5-gram language model with Kneser-Ney smoothing (Kneser and Ney, 1995); lexical estimates of the forward and backward probabilities obtained either by relative frequencies or using the method of (Zens and Ney, 2004); lexicalized distortion (Tillmann, 2004; Koehn et al., 2005); and word count. The parameters of the log-linear model were tuned by optimizing BLEU on the development set using the batch variant of MIRA (Cherry and Foster, 2012). Decoding uses the cube-pruning algorithm of (Huang and Chiang, 2007) with a 7-word distortion limit.

Table 2 shows the performance of these contaminator systems on the test sets used in our

Dataset	Test Segments	Tokens	BLEU	WER
Europarl	2000	59 936	26.05	65.66
EMEA	1832	32 601	31.31	56.90
GALE	1082	33 064	19.65	76.03

Table 2: MT Contaminator Performance

	Baseline		+8× aux.	
Corpus	segments	tokens	segments	tokens
Europarl	222K	6.2M	2.0M	55.5M
EMEA	93K	1.5M	835K	12.6M
GALE	60.6K	2.0M	545K	17.7M

Table 3: Sizes of training datasets: "*Baseline*" contains clean (uncontaminated) data only; in addition, "+8× *aux.*" data also contains contaminated data totalling 8 times the size of the baseline data.

experiments. Using these MT systems, we built sets of training data containing varying degrees of MT contamination, ranging from 0.00 (uncontaminated – or clean – data) to 1.00 (data in which all the target language text is in fact MT output).

As much as possible, when generating contaminated training data, we tried to mimick a realistic situation, in which whole documents, rather than individual segments, are either clean or dirty. This is important, because MT systems, just like human translators, are known to translate terms relatively consistently within documents (Carpuat and Simard, 2012); and because all occurrences of a rare term tend to "bunch up" (Church and Gale, 1995), an MT system is likely to learn its translation from a small number of documents, sometimes a single one. In the case of the EMEA data, since we had access to the document structure, simulating this effect was straightforward. For GALE and Europarl data, we had to settle for an approximation: the original order of segments was preserved, and data was arbitrarily segmented into blocks of 50 contiguous sentences (in other words: we assumed 50-sentence documents).

Contamination was always produced by translating training data in the same direction as the system which it was used to train; for example, when building MT systems to translate from Chinese to English, we used training data containing Chinese source segments paired with their English machine translation. We speculate that this sort of contamination is more likely to have an adverse effect on MT performance than the opposite, and so our setup corresponds to a worst case scenario.[6]

3 Results

We built multiple MT systems for each dataset, starting with a baseline system trained exclusively with clean data, then gradually adding contaminated auxiliary data. Table 3 gives the sizes of the smallest and largest training sets for each test domain.

The "machine-learners" used in all experiments were also built using the Portage SMT toolkit, and a setup similar to that of the English-French contaminator MT systems (Section 2.3). The main difference is that here, we used 4-gram language models instead of 5-gram, which are more appropriate with smaller amounts of training data.

Figure 1 shows learning curves for the MT systems trained with these datasets. The size of the auxiliary data is expressed relative to the size of the baseline training data: 0 means "no

[6] Note that this is in line with Lambert et al. (2011)'s suggestion that, when using MT as complementary training data, it is better to use target-language corpora machine-translated into the source-language, rather than the opposite.

auxiliary data" (baseline conditions), 0.5 means "auxiliary data is half the size of baseline data", etc.

With clean data (black curves), system performance increases more-or-less monotonically as more training data is used, as is usually the case with SMT systems. Depending on the text domain and the initial state of the system, this increase can be quite modest (Europarl) or surprisingly large (EMEA). But in all cases, the behavior is essentially the same. As the degree of MT contamination increases, however, the gains in performance diminish. This reduction is barely visible – and in some cases not statistically significant – with 5% contamination (red curves), but becomes more apparent as the level of contamination increases. In the case of Europarl and EMEA, when the auxiliary training data entirely consists of MT output (100% contamination level – light blue curves), the performance of the resulting system actually decreases with auxiliary data, regardless of the actual amount of data. For Europarl, the contamination threshold at which adding auxiliary training data becomes harmful seems to be somewhere around 0.5: at this level, the performance of the system remains stable as more data is added. For EMEA, this threshold appears to be above 0.5.

In the case of the GALE data, the performance of the system always improves, regardless of the degree of contamination. In this scenario, MT contamination was produced using a rule-based system (SYSTRAN). Complementarity effects between rule-based and statistical systems have been observed in the past, although in the very different context of automatic post-editing (Simard et al., 2007; Dugast et al., 2007), which could explain this difference in behavior. However, another important difference between the Europarl and EMEA scenarios on the one hand, and the GALE scenario on the other hand, is that in the latter, the performance of the baseline system (without auxiliary data; 14.02 BLEU) lies clearly below that of the MT contaminators themselves on the same data (19.65 BLEU; see Table 2). We conjecture that, as long as the quality of the translation in the training data is better than that of the system under training, performance can only increase.

To verify this hypothesis, we compare the learning behavior of systems augmented with 100% contaminated auxiliary data, under different baseline conditions. Figure 2 shows learning curves for systems trained with Europarl and EMEA data[7]; each curve corresponds to a different baseline system.

In the case of Europarl data, the smallest baseline system was trained with only 12.5K clean sentence pairs: at 22.4 BLEU, its performance is below that of the MT contaminators (26.1 BLEU – dotted line). That system always benefits from more auxiliary data, eventually surpassing the MT contaminators. At the opposite end of the spectrum, a Europarl baseline system trained with 200K segments of clean data, with 29.5 BLEU, can only lose from the addition of 100% contaminated data. Intermediate systems behave in between these two extremes. An interesting case is the 50K baseline Europarl system, which initially loses 1 BLEU from doubling its training set with contaminated data, but eventually regains it, as more contaminated data is added. It is worth noting that all systems eventually perform better than the MT contaminators, which is likely explained by the presence of clean in-domain data in the training set (baseline system training data).

A somewhat similar pattern can be observed with EMEA systems. Interestingly in this case, with large amounts of auxiliary training data, all systems converge to almost identical performances, very close to that of the MT contaminator's. In essence, this suggests that all of these systems eventually mimick the behavior of the MT contaminators. This result is in line with those of Dugast et al. (2008), who "relearn" a rule-based MT system with an SMT system by using the former system to generate training data for the latter.

[7] The GALE dataset is too small to produce reliable learning curves with baseline systems performing above the MT contaminator level.

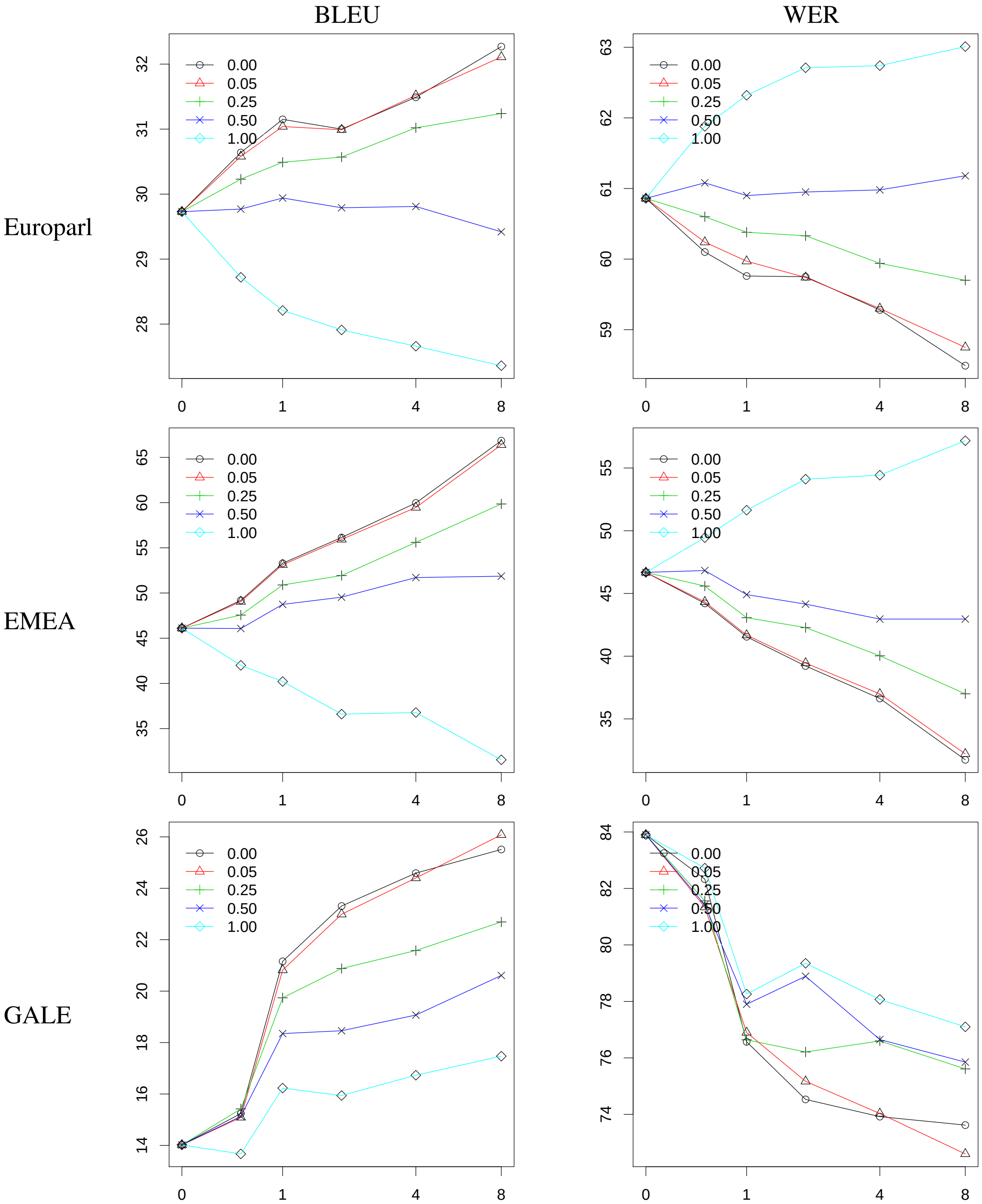

Figure 1: BLEU and WER scores of trained MT system as a function of auxiliary training data size (expressed as a factor of baseline data size); each curve corresponds to a different level of MT contamination in auxiliary data.

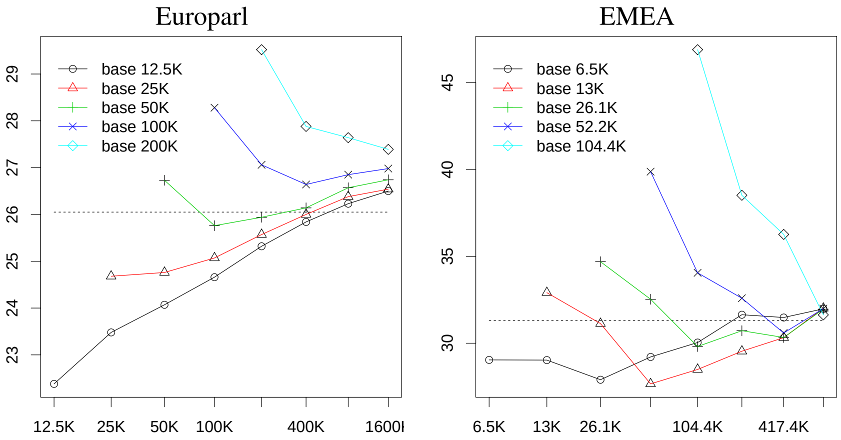

Figure 2: Learning Curves (BLEU score as a function of total amount of training data: baseline + auxiliary) for Europarl and EMEA domains, using 100% contaminated auxiliary data; each curve corresponds to a different amount of (clean) baseline training data. Dotted line is BLEU score of corresponding MT contaminator system on test data.

4 Decontaminating Training Data

Because "dirty" data takes many different forms, data-cleaning procedures are varied. In the case of extra-textual objects or encoding issues, it typically takes the form of in-segment filtering or case-by-case normalization; for misalignment, re-alignment may be necessary; for "bad" translations, re-translation is sometimes considered (Asia Online, 2009). In all cases, filtering out whole segments or even complete documents is usually the simplest solution. And in the case of MT contamination, filtering out dubious portions of the dataset is possibly the only reasonable option.

However, detecting MT is not an easy task, especially when it may be hiding inside large quantities of data, and if it comes from an arbitrary number of different MT systems. Somers et al. (2006) explore different ways of detecting inappropriate uses of online MT systems by language students in translation assignments, a task which is related to that of detecting MT. But their approach is based on comparing human production with MT output, an approach which would be difficult to apply in our case. As mentioned in the introduction, watermarking has also been considered for detecting MT-contaminated data (Venugopal et al., 2011). Unfortunately, this is not a general solution to MT detection because it only allows one producer of MT (e.g. Google) to recognize output from its own systems. In fact, Venugopal et al. suggest that no generally-applicable automatic method exists to distinguish between human- and machine-generated translations. Kurokawa et al. (2009) propose a method to determine whether a given version of a text is the original or a translation; a similar approach could arguably be used for our problem, given suitable training data. To our knowledge, this has not been done, and no such dataset is available.

But what if we knew how to detect MT? What if it was possible to reliably detect contaminated segments or documents, and filter them out of the auxiliary training data? To examine the potential of such methods, we trained systems with artificially decontaminated datasets. Figure 3 compares the behavior of these systems with those obtained with the corresponding contaminated datasets. Each curve plots the BLEU (WER) gain of a system for a given amount of auxiliary data, as a function of the degree of contamination: solid lines correspond to "raw" (contaminated) data, while dashed lines correspond to the same data after decontamination. Because these are “oracle decontaminations”, the results can be interpreted as an empirical upper bound on the MT quality that could be obtained from an actual “MT detector”.

As can be seen in these plots, when the contamination level of the auxiliary data is 0 (left edge of the graphs), filtering has no effect, and thus both systems produce indentical results. When all auxiliary data is MT (right edge of the graphs), filtering eliminates all auxiliary data, and filtered systems display a gain of 0 (black horizontal line); in the case of EMEA and Europarl, the corresponding unfiltered systems display negative gains (i.e. losses), while the GALE systems retain positive gains, as observed earlier. In between these two extremes, the effect of decontamination is generally positive (although there are exceptions). Performance varies as a function of contamination level: the greater the contamination, the greater the effect of decontamination; and as a function of the size of the auxiliary dataset: the greater the amount of data, the greater the effect of decontamination. Yet, it is striking that in most “realistic” scenarios (MT contamination below 20%) the net effect of decontamination is negligible. This suggests that, if the goal is to improve the general quality of the output MT, it is probably not worth investing heavily in developing an MT detector for the specific purpose of cleaning up SMT training data.

5 Discussion

Should one worry about MT contamination when collecting auxiliary training data for SMT systems? It is difficult to answer this question based solely on the results of the experiments reported in the previous sections. One problem is that we have no idea what the actual level of MT contamination is on the Web. In practice, this is likely to depend on various application-dependent factors: where and when the data was collected, what domain and genre, and for which language pair. For some domains and languages, for example English-French legal texts, contamination may be expected to be negligible. For others, for example Russian-English product reviews and descriptions, it may be much higher. In the absence of general MT detection methods, the only way to reliably estimate the level of contamination is to analyse random samples of the collected data manually.

In our experiments, using fully machine translated (i.e. 100% contaminated) auxiliary training data most often degrades MT performance. This appears to contradict earlier results (Ueffing, 2006; Bertoldi and Federico, 2009; Schwenk and Senellart, 2009; Lambert et al., 2011), in which machine translated training data was successfully used to improve MT. The main difference with our work is that all these previous studies were concerned with domain-adaptation: the test and baseline training data were assumed to come from different domains, while the MT auxiliary data came from the same domain as the test data. In our case, all data – test, baseline and auxiliary training – come from the same domain. Again, our setup reflects a worst-case scenario, in which the auxiliary training data is not assumed to be closer to the test than the baseline training data, and our results must be interpreted accordingly.

Another important factor to consider when using potentially contaminated data is the intended application for the trained MT systems. To better understand this point, it is useful to examime some example translations. Table 4 compares example translations for two of our experimental systems on the EMEA data: the *Baseline* system, and the system trained with all available auxiliary training data, under the assumption that this was 100% contaminated (all translations are MT). Example 1 shows how natural “translation memory effects”, which can be normally observed when a domain is by nature very repetitive (as the EMEA), can be lost with contaminated data: a frequent sentence, for which the baseline system “knew” a correct translation, is now translated differently by the augmented system. To a certain extent, this phenomenon may be exacerbated in our study by our experimental methodology: in practice, finding a complete sentence verbatim in a Web-harvested corpus is probably quite unusual.

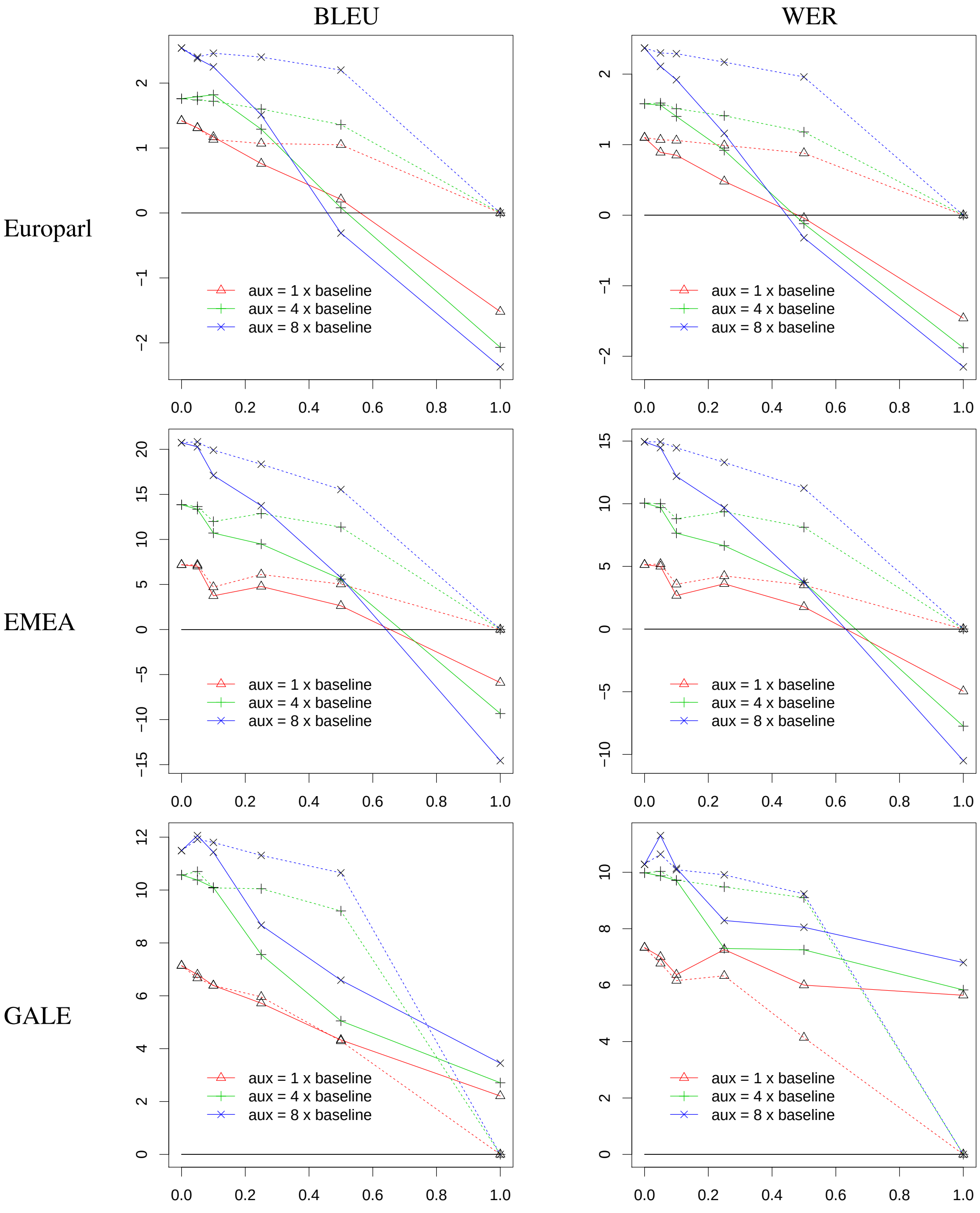

Figure 3: BLEU and WER gain relative to baseline, as a function of MT contamination level in auxiliary training data. Each pair of solid and dashed lines corresponds to a different amount of auxiliary training data (1, 4 and 8 times the size of the baseline dataset): solid lines correspond to contaminated data; dashed lines correspond to filtered (decontaminated) data.

	Source	Reference	Baseline	+ 100% aux.
1.	pour une description complète des effets indésirables observés sous viagra , voir la notice .	for the full list of all side effects reported with viagra , see the package leaflet .	for a full list of all side effects reported with viagra , see the package leaflet .	for a complete description of adverse effects observed under viagra , see the notice .
2.	par la suite , la dose doit être ajustée au cas par cas toutes les 1 à 2 semaine(s) pour maintenir le taux moyen de neutrophiles entre 1,5 x 109/ l et 10 x 109/ l.	subsequently the dose may be individually adjusted every 1 - 2 weeks to maintain the average neutrophil count between 1.5 x 109/ l and 10 x 109/ l.	thereafter , the dose should be adjusted in case by cases every 1 to 2 semaine(s) to maintain the average neutrophil count between 1.5 x 109/ l and 10 x 109/ l.	subsequently , the dose must be adjusted on a case-by-case basis every 1-2 semaine(s) to maintain the average rate of neutrophils between 1.5 x 109/ l and 10 x 109/ l.

Table 4: Translation examples from the **EMEA** test domain. *Baseline* translations are produced by the baseline MT system (see Table 3); *+ 100% aux.* translations are produced by a system trained with all available auxiliary training data (8 times the baseline data), 100% contaminated.

	Source	Reference	Baseline	+ 100% aux.
1.	next , there is - and once again it is inevitable , yet questionable - an element of randomness , which is understandable , in the decisions made by the european institute of florence .	ensuite , il y a , là encore , c' est inévitable , mais néanmoins discutable , une part d' arbitraire , on le comprend , dans les choix opérés par l' institut de florence .	ensuite , il y a une fois de plus - et c' est inévitable , encore contestable - un élément de randomness , ce qui est compréhensible , dans les décisions prises par l' institut de florence .	ensuite , il y a - et , encore une fois , il est inévitable , mais discutable - un élément du caractère aléatoire , ce qui est compréhensible , dans les décisions prises par l' institut européen de florence .
2.	i should be grateful if the court would issue its opinion on the reform of the financial regulation soon .	je remercie la cour des comptes de bien vouloir émettre rapidement son avis sur la réforme du règlement financier .	je serais reconnaissant la cour serait question son avis sur la réforme du règlement financier bientôt .	je devrais être reconnaissants si le tribunal devrait émettre son opinion sur la réforme de la réglementation financière bientôt .

Table 5: Translation examples from the **Europarl** test domain. See Table 4 for details.

But the same example also highlights a more frequent problem: the augmented system may "unlearn" official terminology and standard phraseology from the baseline corpus. For example, the term "*side effect*" which becomes "*adverse effects*", or "*package leaflet*" which is now translated as "*notice*". In Example 2, "*average neutrophil count*" becomes "*average rate of neutrophils*". If the domain is very sensitive to terminology and terminological consistency, this may be a serious problem.

If, however, the emphasis is more on fluency and getting the meaning right, access to more data may turn out to be beneficial, possibly more so than the BLEU and WER scores would have us believe. Examples from the Europarl domain (Table 5) show situations where the translation from the augmented system, although not as close to the reference as the baseline, is actually more fluent and adequate. This seems to be due partly to a better coverage of the source vocabulary (e.g. *randomness*) and to better handling of idioms.

6 Conclusion

We have presented a study on the effect of MT contamination in training data on SMT performance. Our focus was on scenarios where an existing baseline MT system is augmented with Web-collected data, which may contain arbitrary quantities of machine-translated contents. Our experiments demonstrate that MT quality is systematically affected by contaminated training data; severely contaminated data can even decrease the quality of translations compared to the baseline. In all cases of mild contamination, however, the adverse effects are usually quite small, unless very large amounts of data are involved, or in the case of highly technical and/or repetitive material.

Automatic methods to reliably distinguish between human and machine translations would potentially be useful for decontaminating Web-collected data. However, it is uncertain that such methods can be deployed in practice. In the presence of potentially contaminated data, an alternative approach might be to identify which part of the collected training data is more likely to be useful. For example, one could easily identify pairs of segments that contain previously unseen vocabulary. Similarly, one could filter out segments or phrase-pairs that contain domain-specific terminology, or unusual or inconsistent translations of this terminology.

Yet another approach with less-reliable training data is to turn to standard domain-adaptation techniques. For example, auxiliary data could be used to generate distinct phrase tables and language models, whose relative weight in the trained systems would be determined empirically based on performance on suitably chosen tuning sets (Foster and Kuhn, 2007). This is something we plan to integrate into our experimental framework in the near future.

When considering using bilingual data automatically harvested from the Web as training data for MT systems, it is important to take the final application into consideration: whether the resulting MT is intended for post-editing terminology-sensitive material, or for gisting and knowledge acquisition. In the end, the best advice is probably to sample the data for quality, and more importantly to monitor the quality of the resulting MT systems, both by appropriate use of standard benchmarks and metrics, including human evaluation.

7 Acknowledgements

We wish to thank Jean Senellart and the people at SYSTRAN for providing the Chinese-to-English translations for the GALE corpus.

References

Aikawa, T., Schwartz, L., King, R., Corston-Oliver, M., and Lozano, C. (2007). Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. *Proceedings of the MT Summit XI*, pages 10–14.

Asia Online (2009). Study on the Impact of Data Consolidation and Sharing for Statistical Machine Translation, V3.0. Technical report, Asia Online.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further Meta-Evaluation of Machine Translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Carpuat, M. and Simard, M. (2012). The trouble with SMT consistency. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 442–449.

Cherry, C. and Foster, G. (2012). Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada.

Church, K. W. and Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, 1:163–190.

Dugast, L., Senellart, J., and Koehn, P. (2007). Statistical post-editing on systran's rule-based translation system. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 220–223.

Dugast, L., Senellart, J., and Koehn, P. (2008). Can we Relearn an RBMT system? In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 175–178, Columbus, Ohio. Association for Computational Linguistics.

Dyer, C., Gimpel, K., Clark, J. H., and Smith, N. A. (2011). The CMU-ARK German-English Translation System. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 337–343.

Foster, G. and Kuhn, R. (2007). Mixture-Model Adaptation for SMT. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic.

Goutte, C., Carpuat, M., and Foster, G. (2012). The Impact of Sentence Alignment Errors on Phrase-Based Machine Translation Performance. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas*, San Diego, USA.

Huang, L. and Chiang, D. (2007). Forest Rescoring: Faster Decoding with Integrated Language Models. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 144–151, Prague, Czech Republic. Association for Computational Linguistics.

Jiang, J., Way, A., and Carson-Berndsen, J. (2010). Lattice score based data cleaning for phrase-based statistical machine translation. In *EAMT 2010 - 14th Annual Conference of the European Association for Machine Translation*, Saint-Raphaël, France.

Kneser, R. and Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 181–184. IEEE.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of The Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand.

Koehn, P., Axelrod, A., Mayne, A. B., Callison-Burch, C., Osborne, M., and Talbot, D. (2005). Edinburgh System Description for the 2005 IWSLT Speech Translation Evaluation. In *Proceedings of IWSLT-2005*, Pittsburgh, PA.

Kurokawa, D., Goutte, C., and Isabelle, P. (2009). Automatic detection of translated text and its impact on machine translation. In *Proceedings of The twelfth Machine Translation Summit (MT Summit XII)*, Ottawa, Canada.

Lambert, P., Schwenk, H., Servan, C., and Abdul-Rauf, S. (2011). Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293.

Larkin, S., Chen, B., Foster, G., Germann, U., Joanis, E., Johnson, H., and Kuhn, R. (2010). Lessons from NRC's Portage system at WMT 2010. In *the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 127–132.

Madnani, N. (2010). *The circle of meaning: From translation to paraphrasing and back*. PhD thesis, University of Maryland, College Park.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Schwenk, H. and Senellart, J. (2009). Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *MT Summit XII*, Ottawa, Canada.

Simard, M., Goutte, C., and Isabelle, P. (2007). Statistical phrase-based post-editing. In *Proceedings of NAACL HLT*, pages 508–515, Rochester, USA.

Somers, H., Gaspari, F., and Niño, A. (2006). Detecting Inappropriate Use of Free Online Machine-Translation by Language Students-A Special Case of Plagiarism Detection. In *11th Annual Conference of the European Association for Machine Translation–Proceedings*, pages 41–48.

Tiedemann, J. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In *Recent Advances in Natural Language Processing*, volume 5, pages 237–248.

Tillmann, C. (2004). A Unigram Orientation Model for Statistical Machine Translation. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA. Association for Computational Linguistics.

Ueffing, N. (2006). Using Monolingual Source-Language Data to Improve MT Performance. In *International Workshop on Spoken Language Translation (IWSLT) 2006*, pages 174–181, Kyoto, Japan.

Venugopal, A., Uszkoreit, J., Talbot, D., Och, F. J., and Ganitkevitch, J. (2011). Watermarking the outputs of structured prediction with an application in statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1363–1372.

Wohlsen, M. (2014). The Next Big Thing You Missed: Why eBay, Not Google, Could Save Automated Translation. Wired online (http://www.wired.com/2014/04/the-next-big-thing-you-missed-how-ebay-can-change-online-translation).

Yang, J., Senellart, J., and Zajac, R. (2003). Systran's Chinese word segmentation. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 180–183.

Zens, R. and Ney, H. (2004). Improvements in Phrase-Based Statistical Machine Translation. In Susan Dumais, D. M. and Roukos, S., editors, *HLT-NAACL 2004: Main Proceedings*, pages 257–264, Boston, Massachusetts, USA. Association for Computational Linguistics.