Improving Statistical Machine Translation by Adapting Translation Models to Translationese

Gennadi Lembersky∗
University of Haifa, Israel

Noam Ordan∗∗
University of Haifa, Israel

Shuly Wintner†
University of Haifa, Israel

Translation models used for statistical machine translation are compiled from parallel corpora that are manually translated. The common assumption is that parallel texts are symmetrical: The direction of translation is deemed irrelevant and is consequently ignored. Much research in Translation Studies indicates that the direction of translation matters, however, as translated language (translationese) has many unique properties. It has already been shown that phrase tables constructed from parallel corpora translated in the same direction as the translation task outperform those constructed from corpora translated in the opposite direction.

We reconfirm that this is indeed the case, but emphasize the importance of also using texts translated in the “wrong” direction. We take advantage of information pertaining to the direction of translation in constructing phrase tables by adapting the translation model to the special properties of translationese. We explore two adaptation techniques: First, we create a mixture model by interpolating phrase tables trained on texts translated in the “right” and the “wrong” directions. The weights for the interpolation are determined by minimizing perplexity. Second, we define entropy-based measures that estimate the correspondence of target-language phrases to translationese, thereby eliminating the need to annotate the parallel corpus with information pertaining to the direction of translation. We show that incorporating these measures as features in the phrase tables of statistical machine translation systems results in consistent, statistically significant improvement in the quality of the translation.
1. Introduction

Much research in translation studies indicates that translated texts have unique characteristics that set them apart from original texts (Toury 1980; Gellerstam 1986; Toury 1995). Known as translationese, translated texts (in any language) constitute a genre, or a dialect, of the target language, which reflects both artifacts of the translation process and traces of the original language from which the texts were translated. Among the better-known properties of translationese are simplification and explicitation (Blum-Kulka and Levenston 1983; Blum-Kulka 1986; Baker 1993): Translated texts tend to be shorter, to have lower type/token ratio, and to use certain discourse markers more frequently than original texts. Interestingly, translated texts are so markedly different from original ones that automatic classification can identify them with very high accuracy (Baroni and Bernardini 2006; van Halteren 2008; Ilisei et al. 2010; Koppel and Ordan 2011).

Contemporary statistical machine translation (SMT) systems use parallel corpora to train translation models that reflect source- and target-language phrase correspondences. Typically, SMT systems ignore the direction of translation of the parallel corpus. Given the unique properties of translationese, which operate asymmetrically from source to target language, it is reasonable to assume that this direction may affect the quality of the translation. Recently, Kurokawa, Goutte, and Isabelle (2009) showed that this is indeed the case. They trained a system to translate between French and English (and vice versa) using a French-translated-to-English parallel corpus, and then an English-translated-to-French one. They find that in translating into French the latter parallel corpus yields better results (in terms of higher BLEU scores), whereas for translating into English it is better to use the former.

Typically, however, parallel corpora are not marked for direction. Therefore, Kurokawa, Goutte, and Isabelle (2009) trained an SVM-based classifier to predict which side of a bi-text is the origin and which one is the translation, and trained a translation model by utilizing only the subset of the corpus that corresponds to the direction of the task.

We use these results as our departure point, but improve them in two major ways. First, we demonstrate that the other subset of the corpus, reflecting translation in the “wrong” direction, is also important for the translation task, and must not be ignored; second, we show that explicit information on the direction of translation of the parallel corpus, whether manually annotated or machine-learned, is not mandatory. This is achieved by casting the problem in the framework of domain adaptation: We use domain-adaptation techniques to direct the SMT system toward producing output that better reflects the properties of translationese. We show that SMT systems adapted to translationese produce better translations than vanilla systems trained on exactly the same resources. We confirm these findings using automatic evaluation metrics, as well as through a qualitative analysis of the results.

After reviewing related work in Section 2, we begin by replicating the results of Kurokawa, Goutte, and Isabelle (2009) in Section 3. We then (Section 4) explain why translation quality improves when the parallel corpus is translated in the “right” direction. We do so by showing that the subset of the corpus that was translated in the direction of the translation task (the “right” direction, henceforth source-to-target, or $S \rightarrow T$) yields phrase tables that are better suited for translation of the original language than the subset translated in the reverse direction (the “wrong” direction, henceforth target-to-source, or $T \rightarrow S$). We use several statistical measures that indicate the better quality of the phrase tables in the former case.
We then show (Section 5) that using the entire parallel corpus, including texts that are translated both in the “right” and in the “wrong” direction, improves the quality of the results. Next, we investigate several ways to improve the translation quality by adapting a translation model to the nature of translationese, thereby making the output of machine translation more similar to actual, human translation. Specifically, we create two phrase tables, one for the $S \rightarrow T$ portion of the corpus, and one for the $T \rightarrow S$ portion, and combine them into a mixture model using perplexity minimization (Sennrich 2012) to set the model weights. We show that this combination significantly outperforms a simple union of the two portions of the parallel corpus.

Furthermore, we show that the direction of translation used for producing the parallel corpus can be approximated by defining several entropy-based measures that correlate well with translationese, and, consequently, with translation quality. We use the entire corpus, create a single, unified phrase table, and then use these measures, and in particular cross-entropy, as a clue for selecting phrase pairs from this table. The benefit of this method is that not only does it improve the translation quality, but it also eliminates the need to directly predict the direction of translation of the parallel corpus.

The main contribution of this work, therefore, is a methodology that improves the quality of SMT by building translation models that are adapted to the nature of translationese.1 To demonstrate the contribution of our methodology, we conduct in Section 6 a thorough analysis of our results, both quantitatively and qualitatively. We show that translations produced by our best-performing system indeed reflect some well-known properties of translationese better than the output of baseline systems. Furthermore, we provide several examples of SMT outputs that demonstrate in what ways our adapted system generates better results.

2. Related Work

Kurokawa, Goutte, and Isabelle (2009) were the first to address the direction of translation in the context of SMT. They found that a translation model based on the $S \rightarrow T$ portion of the parallel corpus results in much better translation quality than a translation model based on the $T \rightarrow S$ portion. We replicate these results here (Section 3), and view them as a baseline. In taking direction into account, we are faced with two major challenges. First, using only the “right” portion of the corpus results in discarding potentially very useful data. In real-world scenarios this can be crucial, because the proportion between the two portions of the corpus can vary greatly. In the Hansard corpus, used by Kurokawa, Goutte, and Isabelle (2009), only 20% of the corpus is $S \rightarrow T$. We show that the $T \rightarrow S$ portion is also important for machine translation and thus should not be discarded. Using information-theoretic measures, and in particular cross-entropy, we gain statistically significant improvements in translation quality beyond the results of Kurokawa, Goutte, and Isabelle (2009). The second challenge is to overcome the need to (manually or automatically) classify parallel corpora according to direction. We face this challenge by using an adaptation technique.

In previous work, we investigated the relations between translationese and machine translation, focusing on the language model (LM) (Lembersky, Ordan, and Wintner

---

1 This article is a revised and much extended version of Lembersky, Ordan, and Wintner (2012a). Extensions include experiments with several language pairs, in both directions; better adaptation techniques; a new mixture model; and a detailed analysis of the results.
We showed that LMs trained on translated texts yield better translation quality than LMs compiled from original texts. We also showed that perplexity is a good discriminator between original and translated texts. Importantly, we convincingly demonstrated that the differences between translated and original texts are indeed due to effects of translationese, and cannot be attributed to the domain or topic of the texts. Whereas that work focused on the product of translation, namely, the language model, the current study focuses on the process of translation, to wit, the translation model.

Our current work is closely related to research in domain-adaptation. In a typical domain adaptation scenario, a system is trained on a large corpus of “general” (out-of-domain) training material, with a small portion of in-domain training texts. In our case, the translation model is trained on a large parallel corpus, of which some (generally unknown) subset is “in-domain” \(S \rightarrow T\), and some other subset is “out-of-domain” \(T \rightarrow S\). Most existing adaptation methods focus on selecting in-domain data from a general domain corpus. In particular, perplexity is used to score the sentences in the general-domain corpus according to an in-domain language model. Gao et al. (2002) and Moore and Lewis (2010) apply this method to language modeling, and Foster, Goutte, and Kuhn (2010) and Axelrod, He, and Gao (2011) apply this method to translation modeling. Moore and Lewis (2010) suggest a slightly different approach, using cross-entropy difference as a ranking function.

Domain adaptation methods are usually applied at the corpus level, whereas we focus on an adaptation of the phrase table used for SMT. In this sense, our work follows Foster, Goutte, and Kuhn (2010), who weigh out-of-domain phrase pairs according to their relevance to the target domain. They use multiple features that help to distinguish between phrase pairs in the general domain and those in the specific domain. We rely on features that are motivated by the findings of translation studies, having established their relevance through a comparative analysis of the phrase tables. In particular, we use measures such as translation model entropy, inspired by Koehn, Birch, and Steinberger (2009). Additionally, we apply the method suggested by Moore and Lewis (2010) using perplexity ratio instead of cross-entropy difference.

Koehn and Schroeder (2007) suggest a method for adaptation of translation models. They pass two phrase tables directly to the decoder using multiple decoding paths. As we show in Section 5, the application of this method to our scenario does not result in a clear contribution, and we are able to show better results using our proposed method.

Finally, Sennrich (2012) proposes perplexity minimization as a way to set the weights for translation model mixture for domain adaptation. We successfully apply this method to the problem of adapting translation models to translationese, gaining statistically significant improvements in translation quality.

3. Baseline Experiments

3.1 Europarl Experiments

The task we focus on in our experiments is translation from French to English (FR-EN) and from English to French (EN-FR). To establish the robustness of our approach, we also conduct experiments with other translation tasks, including German–English (DE-EN), English–German (EN-DE), Italian–English (IT-EN), and English–Italian (EN-IT). Our corpus is Europarl (Koehn 2005), specifically, portions collected over the years 1996–1999 and 2001–2009. This is a large multilingual corpus, containing sentences
translated from several European languages. In most cases the corpus is annotated with
the original language and the name of the speaker. For each language pair we extract
from the multilingual corpus two subsets, corresponding to the original languages in
which the sentences were produced. For example, in the case of FR-EN we extract
from our corpus all sentences produced in French and translated into English, and all
sentences produced in English and translated into French. All sentences are lowercased
and tokenized using Moses (Koehn et al. 2007). Sentences longer than 80 words are
discarded. Table 1 depicts the size of the subsets whose target language is English.

We use each subset to train two phrase-based statistical machine translation (PB-
SMT) systems (Koehn et al. 2007), translating in both directions between the languages
in each language pair. In other words, we train two PB-SMTs for each translation task,
each based on a parallel corpus produced and translated in a different direction. We
use GIZA++ (Och and Ney 2000) with grow-diag-final alignment, and extract phrases
of length up to 10 words. We prune the resulting phrase tables as in Johnson et al.
(2007), using at most 30 translations per source phrase and discarding singleton phrase
pairs.

We use all Europarl corpora between the years 1996–1999 and 2001–2009 to con-
struct English, German, French, and Italian 5-gram language models, using interpolated
modified Kneser-Ney discounting (Chen 1998) and no cut-off on all n-grams. We use a
specific symbol to mark out-of-vocabulary words (OOVs). The OOV rate is low, less
than 0.5%, and very similar in all our experiments. We use the portion of Europarl
collected over the year 2000 for tuning and evaluation. For each translation task we ran-
domly extract 1,000 parallel sentences for the tuning set and another set of 5,000 parallel
sentences for evaluation. These sentences are originally written in the translation task’s
source language and are translated into the translation task’s target language (in real-
world scenarios, the directionality of the test set is typically known). We use the MERT
algorithm (Och 2003) for tuning and BLEU (Papineni et al. 2002) as our evaluation
metric. We test the statistical significance of the differences between the results using
the bootstrap resampling method (Koehn 2004).

A word on notation: We use $S \rightarrow T$ when the translation direction of the parallel
corpus corresponds to the translation task and $T \rightarrow S$ when a corpus is translated in the
opposite direction to the translation task. For example, suppose the translation tasks are
English-to-French (E2F) and French-to-English (F2E). We use $S \rightarrow T$ when the French-
original corpus is used for the F2E task or when the English-original corpus is used for
the E2F task; and $T \rightarrow S$ when the French-original corpus is used for the E2F task or
when the English-original corpus is used for the F2E task.

Table 2 depicts the BLEU scores of the SMT systems. The data are consistent with
the findings of Kurokawa, Goutte, and Isabelle (2009): Systems trained on $S \rightarrow T$ parallel

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europarl corpus size, in sentences and tokens.</td>
</tr>
<tr>
<td>Original language</td>
</tr>
<tr>
<td>FR-EN</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>DE-EN</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>IT-EN</td>
</tr>
</tbody>
</table>

1003
texts always outperform systems trained on $T \rightarrow S$ texts. The difference in BLEU score can be as high as 3 points.

### 3.2 Hansard Experiments

The corpora used in the Europarl experiments are small (up to 200,000 sentences). Also, the ratio between $S \rightarrow T$ and $T \rightarrow S$ materials varies greatly for different language pairs. To mitigate these issues we use the Hansard corpus, containing transcripts of the Canadian parliament from 1996–2007, as another source of parallel data. The Hansard is a bilingual French–English corpus comprising approximately 80% English-original texts and 20% French-original texts. Crucially, each sentence pair in the corpus is annotated with the direction of translation.

To address the effect of corpus size, we compile six subsets of different sizes (250K, 500K, 750K, 1M, 1.25M, and 1.5M parallel sentences) from each portion (English-original and French-original) of the corpus. Additionally, we use the dev/test section of the Hansard corpus to randomly select French-original and English-original sentences that are used for tuning (1,000 sentences each) and evaluation (5,000 sentences each).

On these corpora we train twelve French-to-English and twelve English-to-French PB-SMT systems using the Moses toolkit (Koehn et al. 2007). We use the same GIZA++ configuration and phrase table pruning as in the Europarl experiments. We also reuse the English and French language models. French-to-English MT systems are tuned and tested on French-original sentences and English-to-French systems on English-original ones.

Table 3 depicts the BLEU scores of the Hansard systems. The data are consistent with our previous findings: Systems trained on $S \rightarrow T$ parallel texts always outperform...
systems trained on $T \rightarrow S$ texts, even when the latter are much larger. For example, a French-to-English SMT system trained on 250,000 $S \rightarrow T$ sentences outperforms a system trained on 1,500,000 $T \rightarrow S$ sentences.

4. Phrase Tables Reflect Facets of Translationese

The baseline results suggest that $S \rightarrow T$ and $T \rightarrow S$ phrase tables differ substantially, presumably due to the different characteristics of original and translated texts. In this section we explain the better translation quality in terms of the better quality of the respective phrase tables, as defined by a number of statistical measures. We first relate these measures to the unique properties of translationese.

Translated texts tend to be simpler than original ones along a number of criteria. Generally, translated texts are not as rich and variable as original ones, and, in particular, their type/token ratio is lower. Consequently, we expect $S \rightarrow T$ phrase tables (which are based on a parallel corpus whose source is original texts, and whose target is translationese) to have more unique source phrases and a lower number of translations per source phrase. A large number of unique source phrases suggests better coverage of the source text, whereas a small number of translations per source phrase means a lower phrase table entropy.

These expectations are confirmed, as the data in Table 4 show. We report the total size of the phrase table in tokens (Total), the number of unique source phrases (Source), and the average number of translations per source phrase (AvgTran), computed on the
24 phrase tables corresponding to our SMT systems. Evidently, \( S \rightarrow T \) phrase tables have more unique source phrases, but fewer translation options per source phrase. This holds uniformly for all 24 tables.

These findings are consistent with our understanding of translationese. Translated texts are not as rich as original ones; their type-to-token ratio is lower, and the variety of syntactic structures is more limited. \( S \rightarrow T \) phrase tables capture correspondences between phrases written in the source language (original) and translated to the target language (translated). Consequently, more different types in the source language correspond to fewer types in the target language. For example, in the FR-EN \( S \rightarrow T \) lexicon trained on 1.5M sentences, the French word réduite (reduced) has 77 translations, whereas in the \( T \rightarrow S \) lexicon the same word has 143 translations. Moreover, in the \( S \rightarrow T \) lexicon the probability of the best translation, reduced, is 41.2%, whereas in the \( T \rightarrow S \) lexicon it is only 28.7%.

A well-established tool for assessing the quality of a phrase table involves entropy-based measures. Phrase table entropy captures the amount of uncertainty involved in choosing candidate translation phrases (Koehn, Birch, and Steinberger 2009). Given a source phrase \( s \) and a phrase table \( T \) with translations \( t \) of \( s \) whose probabilities are \( p(t \mid s) \), the entropy \( H \) of \( s \) is:

\[
H(s) = -\sum_{t \in T} p(t \mid s) \times \log_2 p(t \mid s)
\]

(1)

To compute the phrase table entropy, Koehn, Birch, and Steinberger (2009) search through all possible segmentations of the source sentence to find the optimal covering set of test sentences that minimizes the average entropy of the source phrases in the covering set. We refer to this measure as covering set entropy, or \( \text{CovEnt} \).

We also propose a metric that assesses the quality of the source side of a phrase table. This metric finds the minimal covering set of a given text in the source language using source phrases from a particular phrase table, and outputs the average length of a phrase in the covering set. This measure is referred to as covering set average length, or \( \text{CovLen} \).

Lembersky, Ordan, and Wintner (2011) show that perplexity distinguishes well between translated and original texts. Moreover, perplexity can reflect the degree of “relatedness” of a given phrase to original language or to translationese. Motivated by this observation, we design a cross-entropy-based measure that assesses how well each phrase table fits the properties of translationese. We then build language models from translated texts, and compute the cross-entropy of each target phrase in the phrase tables according to these language models.

Given a language model \( L \), the cross-entropy of a text \( w = w_1, w_2, \ldots w_N \) is:

\[
H(w, L) = -\frac{1}{N} \sum_{i=1}^{N} \log_2 L(w_i)
\]

(2)

Ideally, we would like to use only Hansard data for the language model, but as we already used much of the Hansard data for training the translation model, we use instead an adaptation of an external corpus (Europarl) to the Hansard domain. We...
build language models of translated texts as follows. For English translationese, we extract 170,000 French-original sentences from the English portion of Europarl, and 3,000 English-translated-from-French sentences from the Hansard corpus (disjoint from the training, development, and test sets, of course). We use each corpus to train a trigram language model with interpolated modified Kneser-Ney discounting and no cut-off. All OOV words are mapped to a special token, $\langle unk \rangle$. Then, we interpolate the Hansard and Europarl language models to minimize the perplexity of the target side of the development set ($\lambda = 0.58$, the mixture weight of the Hansard corpus). For French translationese, we use 270,000 sentences from Europarl and 3,000 sentences from Hansard, $\lambda = 0.81$.

Similarly to covering set entropy, covering set cross-entropy (CovCrEnt) finds the optimal covering set of test sentences that minimizes the weighted cross-entropy of the source phrase in the covering set. Given a phrase table $T$ and a language model $L$, the weighted cross-entropy $W$ for a source phrase $s$ is:

$$W(s, L) = -\sum_{t \in T} H(t, L) \times p(t | s)$$

where $H(t, L)$ is the cross-entropy of $t$ according to a language model $L$.

Table 5 lists the entropy-based measures computed on our 24 phrase tables. Again, the data meet our expectations: $S \rightarrow T$ phrase tables uniformly and unexceptionally have lower entropy and cross-entropy, but higher covering set length.

So far, we have established the hypothesis that $S \rightarrow T$ phrase tables better reflect the properties of translationese than $T \rightarrow S$ ones. But does this necessarily affect the quality
of the generated translations? To verify that, we measure the correlation between the quality of the translation, as measured by BLEU (Table 3), with each of the entropy-based metrics. We compute the correlation coefficient $R^2$ (the square of Pearson’s product-moment correlation coefficient) by fitting a simple linear regression model. Table 6 lists the results; clearly, all three measures are strongly correlated with translation quality. Consequently, we use these measures as indicators of better translations, more similar to translationese. Crucially, these measures are computed directly on the phrase table, and do not require reference translations or meta-information pertaining to the direction of translation of the parallel phrase.

### 5. Adaptation of the Translation Model to Translationese

We have thus established the fact that $S \rightarrow T$ phrase tables have an advantage over $T \rightarrow S$ ones that stems directly from the different characteristics of original and translated texts. We have also identified three statistical measures that explain most of the variability in translation quality. We now explore ways for taking advantage of the entire parallel corpus, including translations in both directions, in light of these findings. Our goal is to establish the best method to address the issue of different translation direction components in the parallel corpus.

#### 5.1 Baseline

As a simple baseline we take the union of the two subsets of the parallel corpus. This gives the decoder an opportunity to select phrases from either subset of the corpus, and MERT can be expected to optimize this selection process. For each translation task in Section 3.1, we concatenate the $S \rightarrow T$ and the $T \rightarrow S$ subsets of the parallel corpora and use the union to train an SMT system (henceforth UNION). We use the same language and reordering models, Moses configuration, and the same tuning and evaluation sets as in Section 3.1. Table 7 reports the results. The UNION systems, which

<table>
<thead>
<tr>
<th>System</th>
<th>FR-EN</th>
<th>EN-FR</th>
<th>DE-EN</th>
<th>EN-DE</th>
<th>IT-EN</th>
<th>EN-IT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow T$</td>
<td>33.64</td>
<td>32.11</td>
<td>26.53</td>
<td>16.96</td>
<td>28.70</td>
<td>23.81</td>
</tr>
<tr>
<td>UNION</td>
<td>33.79</td>
<td>32.24</td>
<td>26.76</td>
<td>17.36</td>
<td>29.12</td>
<td>23.70</td>
</tr>
<tr>
<td>MULTI-PATH</td>
<td>33.81</td>
<td>31.95</td>
<td>26.68</td>
<td>17.39</td>
<td>29.11</td>
<td>23.80</td>
</tr>
<tr>
<td>PPLMIN-1</td>
<td>33.86</td>
<td>32.47</td>
<td>26.83</td>
<td>17.80</td>
<td>29.23</td>
<td>23.86</td>
</tr>
<tr>
<td>PPLMIN-2</td>
<td>33.95</td>
<td>32.65</td>
<td>26.77</td>
<td>17.65</td>
<td>29.44</td>
<td>24.01</td>
</tr>
</tbody>
</table>
use twice as much training data as the $S \rightarrow T$ systems, outperform the $S \rightarrow T$ systems for all language pairs except English-to-Italian. Only in three cases out of six (German-to-English, English-to-German, and Italian-to-English), however, is the gain statistically significant. Nevertheless, this indicates that the $T \rightarrow S$ subset contains useful material that can (and does) contribute to translation quality.

We now look at ways to better utilize this portion. First, we train SMT systems with two phrase tables using multiple decoding paths, and combine them in a log-linear model, following Koehn and Schroeder (2007). The performance of this approach (referred to as MULTI-PATH) is either lower or only slightly better than that of the UNION systems (Table 7).

### 5.2 Perplexity Minimization

Next, we look at a linear interpolation of the translation models. We need a way to tune the weights of the translation model components, and we use perplexity minimization, following Sennrich (2012).

Given $n$ phrase tables, we are looking for a set of $n$ weights $\lambda = \lambda_1, \ldots, \lambda_n$, such that $\sum_{i=1}^{n} \lambda_i = 1$, where $\lambda_i$ is the interpolation weight of phrase table $i$. Then, given a phrase pair $(s, t)$, the linear interpolation of the $n$ models is given by:

$$p(s \mid t; \lambda) = \sum_{i=1}^{n} \lambda_i p_i(s \mid t)$$  \hspace{1cm} (4)

To adapt an interpolated translation model to a specific (development) corpus, let $\hat{p}(s, t)$ be the observed, empirical probability of the pair $(s, t)$ in the development corpus. This is obtained by training a phrase table on the development corpus using the standard methodology; the probability of the pair $(s, t)$ is then extracted from the phrase table. The cross entropy $H$ of a translation model with probabilities $p$ to a development corpus with probabilities $\hat{p}$ is defined as:

$$H = - \sum_{s,t} \hat{p}(s, t) \times \log p(s \mid t)$$  \hspace{1cm} (5)

To minimize the cross entropy, we look for a weight vector $\hat{\lambda}$ such that:

$$\hat{\lambda} = \arg \min_{\lambda} - \sum_{s,t} \hat{p}(s, t) \times \log \left( \sum_{i=1}^{n} \lambda_i p_i(s \mid t) \right)$$  \hspace{1cm} (6)

Each feature of the standard SMT translation model (the phrase translation probabilities $p(t \mid s)$ and $p(s \mid t)$, and the lexical weights $\text{lex}(t \mid s)$ and $\text{lex}(s \mid t)$) is optimized independently.

This technique is particularly appealing for us due to two reasons: first, Lembersky, Ordan, and Wintner (2011) show that perplexity is a good differentiator between original and translated texts; second, the perplexity is minimized with respect to some development set. Consequently, if we use a $S \rightarrow T$ corpus for this purpose, we directly adapt the interpolated phrase table to the qualities of the $S \rightarrow T$ translation models as described in Section 4. We use this technique to interpolate two pairs of phrase tables: we interpolate the $S \rightarrow T$ and the $T \rightarrow S$ models (we refer to this system as
PPLMIN-1) and we also interpolate the $S \rightarrow T$ with the UNION models (PPLMIN-2), as a simple way of upweighting. Table 7 reports the results. In all cases, the interpolated systems yield higher BLEU scores than the simple UNION systems. Although the improvements are small (0.2–0.4 BLEU points), they are statistically significant in all cases, except for German-English. Clearly, the interpolated systems outperform the $S \rightarrow T$ systems by 0.2-0.7 BLEU points (statistically significant in all cases). PPLMIN-2 seems to be better than PPLMIN-1 in four out of six systems.

To verify that the improvement in translation quality is due to the adaptation process rather than a quirk resulting from MERT instability, we use MultEval (Clark et al. 2011). This is a script that takes machine translation hypotheses from several (in our case, three) runs of an optimizer (MERT) and reports three popular metric scores: BLEU, Meteor (Denkowski and Lavie 2011), and TER (Snover et al. 2006). Meteor and BLEU scores are higher for better translations (↑), whereas TER is a lower-is-better measure (↓). In addition, MultEval computes the ratio of output length to reference length (closer to 100% is better), as well as p-values (via approximate randomization). We use MultEval to compare translation hypotheses of the UNION and PPLMIN-2 systems. Table 8 presents the results for French-to-English and English-to-French (other translation tasks produce similar results). The improvement of the adapted systems is clear and robust.

5.3 Adaptation without Explicit Information on Directionality

A prerequisite for interpolating translation models, the method we advocate here, is that the direction of translation of every sentence pair in the parallel corpus be known in advance. When such information is not available, machine learning can automatically classify texts as original or translated (Baroni and Bernardini 2006; van Halteren 

<table>
<thead>
<tr>
<th>Table 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultEval scores for UNION and PPLMIN-2 systems.</td>
</tr>
<tr>
<td>Metric</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td><strong>French-to-English</strong></td>
</tr>
<tr>
<td>BLEU ↑</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>METEOR ↑</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>TER ↓</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>English-to-French</strong></td>
</tr>
<tr>
<td>BLEU ↑</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>METEOR ↑</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>TER ↓</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Length</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Naturally, however, the quality of the interpolation of translation models trained on classified (rather than annotated) data is expected to decrease. In this section we establish an adaptation technique that does not rely on explicit information pertaining to the direction of translation, but rather uses perplexity-based measures to evaluate the “relatedness” of a specific phrase to an original or a translated language “dialect.”

For the following experiments we use the Hansard corpus described in Section 3.2; FO (French original) refers to subsets of the parallel corpus that were translated from French to English, EO (English original) refers to texts translated from English to French. We create three different mixtures of FO and EO: a balanced mix comprising 500K sentences each of FO and EO (MIX), an EO-biased mix with 500K sentences of FO and 1M sentences of EO (MIX-EO), and an FO-biased mix with 1M sentences of FO and 500K sentences of EO (MIX-FO). We use these corpora to train French-to-English and English-to-French MT systems, evaluating their quality on the evaluation sets described in Section 3.2. We use the same Moses configuration as well as the same language and reordering models as in Section 3.2.

Now, we adapt the translation models by adding to each phrase pair in the phrase tables an additional factor, as a measure of its fitness to the genre of translationese. The factors are used as additional features in the phrase table. We experiment with two such factors. First, we use the language models described in Section 4 to compute the cross-entropy of each translation option according to this model. We add cross-entropy as an additional score of a translation pair that can be tuned by MERT (we refer to this system as CrEnt). Because cross-entropy is a “the lower the better” metric, we adjust the range of values used by MERT for this score to be negative.

Second, following Moore and Lewis (2010), we define an adapting feature that not only measures how close phrases are to translated language, but also how far they are from original language, and use it as a factor in a phrase table (this system is referred to as PplRatio). We build two additional language models of original texts as follows. For original English, we extract 135,000 English-original sentences from the English portion of Europarl, and 2,700 English-original sentences from the Hansard corpus. We train a trigram language model with interpolated modified Kneser-Ney discounting on each corpus and we interpolate both models to minimize the perplexity of the source side of the development set for the English-to-French translation task ($\lambda = 0.49$). For original French, we use 110,000 sentences from Europarl and 2,900 sentences from Hansard, $\lambda = 0.61$. Finally, for each target phrase $t$ in the phrase table we compute the ratio of the perplexity of $t$ according to the original language model $L_o$ and the perplexity of $t$ with respect to the translated model $L_t$ (see Section 4). In other words, the factor $F$ is computed as follows:

$$F(t) = \frac{H(t, L_o)}{H(t, L_t)}$$  (7)

We apply these techniques to the French-to-English and English-to-French phrase tables built from the concatenated corpora, and use each phrase table to train an SMT system. We compare the performance of these systems to that of $S \rightarrow T$, UNION, and both PPLMIN systems. Table 9 summarizes the results.

All systems outperform the corresponding UNION systems. CrEnt systems show significant improvements ($p < 0.05$) on balanced scenarios (MIX) and on scenarios biased towards the $S \rightarrow T$ component (MIX-FO in the French-to-English task, MIX-EO in English-to-French). PplRatio systems exhibit more consistent behavior, showing
Table 9
Adaption without classification results.

<table>
<thead>
<tr>
<th>System</th>
<th>MIX</th>
<th>MIX-EO</th>
<th>MIX-FO</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow T$</td>
<td>35.21</td>
<td>35.21</td>
<td>35.73</td>
</tr>
<tr>
<td>UNION</td>
<td>35.27</td>
<td>35.36</td>
<td>35.94</td>
</tr>
<tr>
<td>PPLMIN-1</td>
<td>35.46</td>
<td>35.59</td>
<td>36.26</td>
</tr>
<tr>
<td>PPLMIN-2</td>
<td><strong>35.75</strong></td>
<td>35.65</td>
<td>36.20</td>
</tr>
<tr>
<td>CrEnt</td>
<td>35.54</td>
<td>35.45</td>
<td><strong>36.75</strong></td>
</tr>
<tr>
<td>PplRatio</td>
<td>35.59</td>
<td><strong>35.78</strong></td>
<td>36.22</td>
</tr>
</tbody>
</table>

Task: English-to-French

<table>
<thead>
<tr>
<th>System</th>
<th>MIX</th>
<th>MIX-EO</th>
<th>MIX-EO</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S \rightarrow T$</td>
<td>29.15</td>
<td>29.15</td>
<td>29.94</td>
</tr>
<tr>
<td>UNION</td>
<td>29.27</td>
<td>29.44</td>
<td>30.01</td>
</tr>
<tr>
<td>PPLMIN-1</td>
<td>29.64</td>
<td>29.94</td>
<td>29.65</td>
</tr>
<tr>
<td>PPLMIN-2</td>
<td>29.50</td>
<td><strong>30.45</strong></td>
<td>30.12</td>
</tr>
<tr>
<td>CrEnt</td>
<td>29.47</td>
<td>29.45</td>
<td><strong>30.44</strong></td>
</tr>
<tr>
<td>PplRatio</td>
<td><strong>29.65</strong></td>
<td>29.62</td>
<td>30.34</td>
</tr>
</tbody>
</table>

small, but statistically significant improvement ($p < 0.05$) in all scenarios. Additionally, the new systems perform quite competitively compared to the interpolated systems, winning in four out of six cases. Note again that all systems in the same column (except $S \rightarrow T$) are trained on exactly the same corpus and have exactly the same phrase tables. The only difference is an additional factor in the phrase table that “encourages” the decoder to select translation options that are closer to translated texts than to original ones.

6. Analysis

We have demonstrated that SMT systems that are sensitive to the direction of translation perform better. The superior quality of SMT systems that are adapted to translationese is reflected in higher BLEU scores, but also in the scores of other automatic measures for evaluating the quality of machine translation output. In this section we analyze the better performance of translationese-adapted systems, both quantitatively and qualitatively, relating it to established insights in translation studies.

It may be claimed that translationese-aware systems perform better (in terms of BLEU scores) not because of the properties of translated texts, but due to the closer domain, genre, or topic of translated texts to those of the reference translations. In our previous work (Lembersky, Ordan, and Wintner 2011, 2012b), we convincingly demonstrated that this is not the case, by means of several experiments that abstracted the texts away from specific words. Although these results are concerned with the language model, we trust that they also hold for the translation model on which we focus here.

Furthermore, improvements in BLEU scores that result from attention to translationese are consistent with human judgments. In other words, a machine translation system that produces translations with higher BLEU scores by taking into account
the directionality of translation also produces translations that human judges prefer. Although we have not conducted such experiments here, we have shown this correlation in a previous work (Lembersky, Ordan, and Wintner 2012b) that focused on the language model rather than on the translation model.

6.1 Quantitative Analysis

Is the output of translationese-adapted systems indeed more similar to translationese? We begin with a set of properties of translationese that are easy to compute, and evaluate the output of our translationese-adapted SMT systems in terms of these properties.

6.1.1 Type–Token Ratio. Translated texts have been shown to have lower type-to-token ratio (TTR) than original ones (Al-Shabab 1996). Figure 1 compares the TTR of the translation outputs of $S \rightarrow T$, $T \rightarrow S$, UNION, and PPLMIN-2 systems. For comparison, we also add the TTR of the reference translations for each task. To mitigate the effects of the different morphological systems of the various languages, we compute the TTR in terms of lemmas, rather than surface forms. Obviously, the TTR of $S \rightarrow T$ output is much lower than $T \rightarrow S$ system. Recall that $S \rightarrow T$ systems produce markedly better translations than $T \rightarrow S$ ones, so indeed there is a clear correspondence between the TTR of the outputs and better translation quality. Figure 1 also compares the TTR of the outputs produced from two combination systems, UNION and PPLMIN-2. The UNION outputs are arbitrary: Their TTR is sometimes lower than the corresponding $S \rightarrow T$ system, but sometimes higher than even the corresponding $T \rightarrow S$ system. In contrast, PPLMIN-2 systems (which are the best adapted systems) systematically produce outputs with the lowest TTR, that is, outputs closest to translationese. As expected, reference translations exhibit the lowest TTR in four out of the six tasks.

![Figure 1](http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00159)

**Figure 1**
Type–token ratio in SMT translation outputs.
6.1.2 Singletons. A related property of translated texts is that they tend to exhibit a much lower rate of words that occur only once in a text (hapax legomena) than original texts. We thus count the number of singletons in the outputs of each of the SMT systems (and, for comparison, the reference translations). The results, which are depicted in Figure 2, are not totally conclusive, but are interesting nonetheless. Specifically, in all cases the PPLMIN-2 system exhibits a lower number of singletons than the UNION system; and in all systems except the English-Italian one, the number of singletons produced by the PPLMIN-2 system is lowest. Reference translations exhibit the lowest rate of singletons in five out of the six tasks.

6.1.3 Entropy. As another quantitative measure of the contribution of perplexity minimization as a method of adaptation, we list in Table 10 the values of the entropy-based measures discussed in Section 4 on three types of SMT systems: those compiled from $S \rightarrow T$ texts only, UNION, and PPLMIN-2 ones. Observe that the covering set cross-entropy measure, designed to reflect the fitting of a phrase table’s target side to translated texts, is significantly lower in PPLMIN-2 systems than in $S \rightarrow T$ and UNION systems. This indicates that perplexity minimization improves the system’s fitness to translationese. Interestingly, the PPLMIN-2 systems have better lexical coverage than the UNION systems. Table 10 lists data for French-English and English-French, but other language pairs exhibit similar behavior.

6.1.4 Mean Occurrence Rate. Original texts are known to be lexically richer than translated ones; in particular, translationese uses more frequent and common words (Laviosa 1998). To assess the lexical diversity of a given text we define Mean Occurrence Rate (MOR). MOR computes the average number of occurrences of tokens in the text with
Table 10
Entropy-based measures, computed on phrase tables of baseline and adapted SMT systems.

<table>
<thead>
<tr>
<th>System</th>
<th>CovEnt</th>
<th>CovCrEnt</th>
<th>CovLen</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR-EN</td>
<td>S → T</td>
<td>0.43</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td>T → S</td>
<td>0.45</td>
<td>2.77</td>
</tr>
<tr>
<td>UNION</td>
<td>0.43</td>
<td>2.20</td>
<td>2.34</td>
</tr>
<tr>
<td>PPLMIN-2</td>
<td>0.43</td>
<td>2.14</td>
<td>2.35</td>
</tr>
<tr>
<td>UNION</td>
<td>0.61</td>
<td>3.09</td>
<td>2.17</td>
</tr>
<tr>
<td>PPLMIN-2</td>
<td>0.61</td>
<td>2.99</td>
<td>2.18</td>
</tr>
</tbody>
</table>

respect to a large reference corpus. Consequently, sentences containing more frequent words have higher MOR scores. More formally, given a reference corpus \( R \) with \( n \) word types \( r_1 \cdots r_n \), let \( C(r_i) \) be a number of occurrences of the word \( r_i \) in the corpus \( R \). Then the MOR of a sentence \( S = s_1 \cdots s_k \) is:

\[
MOR(S) = \frac{1}{k} \sum_{i=1}^{k} \log(C(s_i))
\]

(8)

\( C(s_i) \) is calculated from the corpus \( R \) if \( s_i \in R \). Otherwise, \( C(s_i) = \alpha \), where \( \alpha \) is a pre-defined constant depending on the size of the reference corpus. In all our experiments we use \( \alpha = 0.5 \).

In order to establish the relation between the MOR measure and translation quality, we compute MOR scores for each sentence of an SMT system output. Then, we sort the output sentences based on their MOR scores, split the output into two parts—below and above the median of MOR—and calculate BLEU score for each portion independently. We perform these calculations on the outputs of UNION and PPLMIN-2 SMT systems for all our translation tasks. We use the Europarl corpus (Koehn 2005) as a reference for a list of occurrences. Table 11 depicts the results. In all cases, the bottom part (below the median) of SMT outputs has significantly lower BLEU scores (up to 5 BLEU points!) than the upper part, indicating that the MOR measure is a good (post factum) differentiator between poor and good translations.

Table 11
BLEU scores computed on portions of UNION and PPLMIN-2 systems outputs below and above the MOR median.

<table>
<thead>
<tr>
<th>Task</th>
<th>UNION Bottom</th>
<th>UNION Upper</th>
<th>PPLMIN-2 Bottom</th>
<th>PPLMIN-2 Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>DE-EN</td>
<td>24.05</td>
<td>28.72</td>
<td>24.10</td>
<td>28.71</td>
</tr>
<tr>
<td>EN-DE</td>
<td>16.06</td>
<td>18.78</td>
<td>16.42</td>
<td>18.82</td>
</tr>
<tr>
<td>FR-EN</td>
<td>31.48</td>
<td>35.49</td>
<td>31.85</td>
<td>35.49</td>
</tr>
<tr>
<td>EN-FR</td>
<td>28.97</td>
<td>34.83</td>
<td>29.30</td>
<td>35.58</td>
</tr>
<tr>
<td>IT-EN</td>
<td>26.07</td>
<td>31.75</td>
<td>26.43</td>
<td>31.97</td>
</tr>
</tbody>
</table>
We now compute the average MOR score on the outputs of all our SMT systems. Figure 3 shows the results. In all cases (except Italian to English), $S \rightarrow T$ is better than $T \rightarrow S$; and in all systems except EN-FR, PPLMIN-2 is best.

6.2 Qualitative Analysis

Translation is sometimes described as an attempt to strike a balance between interference, the so-called inevitable marks left by the source language on the target text, and standardization, the attempt of the translator to adapt the translation product to the target language and culture, to break away from the source text towards a more adequate text (Toury 1995). In order to study the effect of the adaptation qualitatively, rather than quantitatively, we focus on several concrete examples. We compare translations produced by the UNION (henceforth baseline) and by the PPLMIN-2 (henceforth adapted) French-English Europarl systems. We selected 200 sentences from the French-English evaluation set for manual inspection, focusing on sentences in which the translations were significantly different from each other. Indeed, we find that the translations are better adapted along several dimensions.

In the following sentences, the baseline follows a more literal translation, whereas the adapted system creates a more adequate, standardized translation.

**Source** Monsieur le président, chers collègues, les tempêtes qui ont ravagé la france dans la nuit des 26 et 27 décembre ont fait, on l’a dit, 90 morts, 75 milliards de francs, soit 11 milliards d’euros, de dégâts.

**Baseline** Mr president, ladies and gentlemen, storms that have ravaged france during the night of 26 and 27 december were, as has been said, 90 people dead, 75 billion francs, that is, eur 11 billion, damage.
Adapted  Mr president, ladies and gentlemen, the storms which have devastated France during the night of 26 and 27 December were, as has been said, 90 people dead, 75 billion francs, or €11 billion, damage.

Source  Tout d’abord, je tiens à saluer tous mes collègues maires, élus locaux, qui, au quotidien, ont dû rassurer la population, organiser la solidarité, coopérer avec les services publics.

Baseline  First of all, I should like to pay tribute to all my colleagues, mayors, local elected representatives, who, in their daily lives, have had to reassure the population, organise solidarity, cooperate with public services.

Adapted  First of all, I should like to pay tribute to all my colleagues, mayors, local elected representatives, who, on a daily basis, have had to reassure the population, organise solidarity, cooperate with public services.

Source  Monsieur le président, je vous remercie de me laisser conclure, et je rappellerai simplement une maxime: “les tueurs en série se font toujours prendre par la police quand ils accélèrent la cadence de leurs crimes”.

Baseline  Mr president, thank you for allowing me to leave conclusion, and I would like to remind you just a maxim: ‘the murderers in series are always take by the police when they accélèrent the pace of their crimes’.

Adapted  Mr president, thank you for letting me finish, and I would like to remind you just a maxim: ‘the murderers in series are always take by the police when they accélèrent the pace of their crimes’.

Note that the baseline is not necessarily incomprehensible, nor even “impossible” in the target language; in the first example, it is clear what is meant by storms that have ravaged France, and moreover, we find such expressions in a 1.5-G token-sized corpus (Ferraresi et al. 2008); it is just half as likely as what is offered by the adapted system.

Another difference between the two systems is reordering. Sometimes, as in the two following examples, the inability of the baseline system to reorder the words correctly stems from interference:

Source  Madame la présidente, mes chers collègues, nous croyions, jusqu’à présent, que l’union européenne était, selon les dispositions des traités de Rome et de Paris qui avaient fondé les communautés, devenues union, une association d’États libres, indépendants et souverains.

Baseline  Madam president, ladies and gentlemen, we croyions, up to now, that the European union is, according to the provisions of the treaties of Rome and Paris who had based the communities, become union, an association of states free, independent and sovereign.

Adapted  Madam president, ladies and gentlemen, we croyions, up to now, that the European union was, according to the provisions of the treaties of Paris and Rome who had based communities, become union, an association of free, sovereign and independent states.

Source  La convention de Lomé bénéficie essentiellement à quelques grands groupes industriels ou financiers qui continuent à piller ces pays et perpétuent leur dépendance économique, notamment des anciennes puissances coloniales.

Baseline  The Lomé convention has mainly to a few large industrial groups or financial which continue to plunder those countries and perpétuent their economic dependence, in particular the former colonial powers.
Adapted The romé convention has mainly to a few large financial and industrial groups which continue to plunder those countries and perpetuate their economic dependence, in particular the former colonial powers.

Additionally, the adapted system produces much better collocations. Compare the “natural” expressions pay a high price and express the concern with the baseline system products:

Source Ces hommes et ces femmes qui bougent à travers l’Europe payent leur voiture, leurs taxes nationales, leur pot catalytique, leurs taxes sur les carburants, et payent donc déjà très cher le prix de la magnifique machine et la liberté de circuler.

Baseline These men and women who are moving across Europe are paying their car, their national taxes, their catalytic converter, their taxes on fuel, and therefore already pay very dearly for the price of the magnificent machine and freedom of movement.

Adapted These men and women who are moving across Europe pay their car, their national taxes, their catalytic converter, their taxes on fuel, and therefore already pay a high price for the magnificent machine and freedom of movement.

Source Je veux dire également le souci que j’ai d’une bonne coopération entre interreg et le Féd, notamment pour les caraïbes et l’océan indien.

Baseline I would like to say to the concern that I have good cooperation between interreg and the edf, particularly for the caribbean and the indian ocean.

Adapted I also wish to express the concern that I have good cooperation between interreg and the edf, particularly for the caribbean and the indian ocean.

Last, there are a few cases of explicitation. Blum-Kulka (1986) observed the tendency of translations to introduce to the target texts cohesive markers in order to render implicit utterances more explicit. Koppel and Ordan (2011), who used function words to discriminate between translated and non-translated texts, found that cohesive markers, words such as in fact, however, moreover, and so forth, were among the top markers of translationese, irrespective of source language and domain. And truly we find them also over-represented in the adapted system:

Source Nous affirmons au contraire la nécessité politique de rééquilibrer les rapports entre l’Afrique et l’Union européenne.

Baseline We say the opposite the political necessity to rebalance relations between africa and the european union.

Adapted on the contrary, we maintain the political necessity of rebalancing relations between africa and the european union.

Source Cette mention semble alors contredire les explications linguistiques données par l’office et laisse craindre que l’erreur ne revête pas le seul caractère technique que l’on semble vouloir lui donner.

Baseline This note seems so contradict the explanations given by the language and leaving office fear that the mistake revête do not only technical nature hat we seems to want to give it.

Adapted This note therefore seems to contradict the linguistic explanations given by the office and fear that leaves the mistake revête do not only technical nature that we seems to want to give it.

In (human) translation circles, translating out of one’s mother tongue is considered unprofessional, even unethical (Beeby 2009). Many professional associations in Europe urge translators to work exclusively into their mother tongue (Pavlović 2007). The two
kinds of automatic systems built in this article reflect only partly the human situation, but they do so in a crucial way. The \( S \rightarrow T \) systems learn examples from many human translators who follow the decree according to which translation should be made into one’s native tongue. The \( T \rightarrow S \) systems are flipped directions of humans’ input and output. The \( S \rightarrow T \) direction proved to be more fluent and accurate. This has to do with the fact that the translators “cover” the source texts more fully, having a better “translation model.”

7. Combining Translation and Language Models

When we experimented with translation models, we compiled language models from corpora comprising original and translated texts. Our previous work (Lembersky, Ordan, and Wintner 2011, 2012b), however, shows that language models compiled from translated texts are better for machine translation than models trained on original texts. In this section we examine whether these findings have a cumulative effect. In other words, we test if an additional improvement in translation quality can be gained by combining our findings for both the language and the translation model.

The tasks are translating French-to-English (FR-EN) and English-to-French (EN-FR). We re-use the Europarl-based translation models of Section 3.1. We compile language models from the French-English Hansard-based parallel corpora described in Section 3.2. We use 1 million parallel sentence subsets. We train an original French LM on the source side of the \( S \rightarrow T \) corpus and we train the translated English LM on the target side of the same corpus. In the same manner we compile the translated French LM and the original English LM from the \( T \rightarrow S \) corpus. All language models are 5-grams with an interpolated modified Kneser-Ney discounting (Chen 1998). The vocabulary is limited to tokens that appear twice or more in the reference set. All unknown words are mapped to a special token. We tune and evaluate all SMT systems on two kinds of reference sets: Europarl (Section 3.1) and Hansard (Section 3.2).

First, we use all possible combinations of translation and language models to train four SMT systems for each translation task: \( T \rightarrow S \) TM with original (O) LM, \( T \rightarrow S \) TM with translated (T) LM, \( S \rightarrow T \) TM with O LM, and \( S \rightarrow T \) TM with T LM. All systems are tuned and evaluated on both the Europarl and the Hansard reference sets. Table 12

<table>
<thead>
<tr>
<th>FR-EN (EUROPARL)</th>
<th>FR-EN (HANSARD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LM</strong></td>
<td><strong>LM</strong></td>
</tr>
<tr>
<td><strong>O</strong></td>
<td><strong>T</strong></td>
</tr>
<tr>
<td><strong>T</strong></td>
<td><strong>O</strong></td>
</tr>
<tr>
<td>( T \rightarrow S )</td>
<td>27.06</td>
</tr>
<tr>
<td>( S \rightarrow T )</td>
<td>30.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EN-FR (EUROPARL)</th>
<th>EN-FR (HANSARD)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LM</strong></td>
<td><strong>LM</strong></td>
</tr>
<tr>
<td><strong>O</strong></td>
<td><strong>T</strong></td>
</tr>
<tr>
<td><strong>T</strong></td>
<td><strong>O</strong></td>
</tr>
<tr>
<td>( T \rightarrow S )</td>
<td>22.33</td>
</tr>
<tr>
<td>( S \rightarrow T )</td>
<td>25.11</td>
</tr>
</tbody>
</table>
shows the translation quality of the SMT systems in terms of BLEU. Both translation and language models contribute to the translation quality, but it seems that the contribution of the translation model is more significant. Even in the case of the Hansard reference set, in the English-to-French translation task, the $S \rightarrow T$ TM (compiled from Europarl texts) adds 1.2 BLEU points, and the $T \rightarrow S$ LM (compiled from Hansard texts) adds only 0.46 BLEU points.

Next, we perform a set of experiments to test whether a combination of the adaptation techniques described in Section 5 for the translation and language models can further improve the translation quality. First, we build a baseline SMT system with a translation model trained on a concatenation of $S \rightarrow T$ and $T \rightarrow S$ parallel corpora and a language model compiled from a concatenation of translated and original texts. Then, we build two other systems, one with an adapted translation model and one with an adapted language model. Finally, we use the adapted translation and language models to train yet another SMT system. We use the PPLMIN-2 method (Section 5) to adapt the translation model and linear interpolation to adapt the language model. The SMT systems are then tuned and evaluated on the Europarl and the Hansard reference sets. The results, depicted in Table 13, show that SMT systems with an adapted TM usually outperform the baseline systems. LM adaptation alone does not improve the translation quality, but if combined with TM adaptation it produces the best results (but not significantly better than just TM adaptation).

**8. Conclusion**

Phrase tables trained on parallel corpora that were translated in the same direction as the translation task perform better than ones trained on corpora translated in the opposite direction. Nonetheless, even “wrong” phrase tables contribute to the translation quality. We analyze both “correct” and “wrong” phrase tables, uncovering a great deal of difference between them. We use insights from translation studies to explain these differences; we then adapt the translation model to the nature of translationese.

We investigate several approaches to the adaptation problem. First, we use linear interpolation to create a mixture model of $S \rightarrow T$ and $T \rightarrow S$ translation models. We
use perplexity minimization and an $S \rightarrow T$ reference set to determine the weights of each model, thus directly adapting the model to the properties of translationese. We show consistent and statistically significant improvements in translation quality on three different language pairs (six translation tasks) using several automatic evaluation metrics.

Furthermore, we incorporate information-theoretic measures that correlate well with translationese into phrase tables as an additional score that can be tuned by MERT, and show a statistically significant improvement in the translation quality over all baseline systems. We also analyze the results qualitatively, showing that SMT systems adapted to translationese tend to produce more coherent and fluent outputs than the baseline systems. An additional advantage of our approach is that it does not require an annotation of the translation direction of the parallel corpus. It is completely generic and can be applied to any language pair, domain, or corpus.

Where this work focuses on improving the process of translation (i.e., the translation model), our previous work (Lembersky, Ordan, and Wintner 2012b) focuses on improving the product of translation (i.e., the language model). We have shown preliminary results in which both models were adapted to translationese; an open challenge is finding the optimal combination of improving both process and product in a single unified system.

**Acknowledgments**

We are grateful to Cyril Goutte, George Foster, and Pierre Isabelle for providing us with an annotated version of the Hansard corpus. Alon Lavie has been instrumental in stimulating some of the ideas reported in this article, as well as in his long-term support and advice. We benefitted greatly from several constructive suggestions by the three anonymous *Computational Linguistics* referees. This research was supported by the Israel Science Foundation (grant no. 137/06) and by a grant from the Israeli Ministry of Science and Technology.

**References**


Lembersky, Gennadi, Noam Orban, and Shuly Wintner. 2012b. Language models...
for machine translation: Original vs.
translated texts. Computational Linguistics,

Moore, Robert C. and William Lewis.
2010. Intelligent selection of language
model training data. In Proceedings of
the ACL 2010 Conference, Short Papers,
pages 220–224, Stroudsburg, PA.

Och, Franz Josef. 2003. Minimum error rate
training in statistical machine translation.
In ACL ‘03: Proceedings of the 41st Annual
Meeting on Association for Computational
Linguistics, pages 160–167, Morristown, NJ.

Improved statistical alignment models.
In ACL ‘00: Proceedings of the 38th Annual
Meeting on Association for Computational
Linguistics, pages 440–447, Morristown, NJ.

Papineni, Kishore, Salim Roukos, Todd
Ward, and Wei-Jing Zhu. 2002. BLEU:
a method for automatic evaluation of
machine translation. In ACL ‘02:
Proceedings of the 40th Annual Meeting on
Association for Computational Linguistics,
pages 311–318, Morristown, NJ.

Pavlović, Nataša. 2007. Directionality in
translation and interpreting practice.
Report on a questionnaire survey in

Sennrich, Rico. 2012. Perplexity
minimization for translation model
domain adaptation in statistical
machine translation. In Proceedings
of the 13th Conference of the European
Chapter of the Association for
Computational Linguistics,
pages 539–549, Avignon.

Snover, Matthew, Bonnie Dorr, Richard
Schwartz, Linnea Micciulla, and John
Makhoul. 2006. A study of translation
edit rate with targeted human annotation.
In Proceedings of the 7th Conference of the
Association for Machine Translation of the
Americas (AMTA-2006), pages 223–231,
Cambridge, MA.

Toury, Gideon. 1980. In Search of a Theory of
Translation. The Porter Institute for Poetics
and Semiotics, Tel Aviv University,
Tel Aviv.

Toury, Gideon. 1995. Descriptive Translation
Studies and Beyond. John Benjamins,
Amsterdam / Philadelphia.

van Halteren, Hans. 2008. Source
language markers in EUROPARL
translations. In COLING ’08: Proceedings
of the 22nd International Conference on
Computational Linguistics, pages 937–944,
Morristown, NJ.