Adapting Translation Models to Translationese Improves SMT

Gennadi Lembersky  
Dept. of Computer Science  
University of Haifa  
31905 Haifa, Israel  
glembers@campus.haifa.ac.il

Noam Ordan  
Dept. of Computer Science  
University of Haifa  
31905 Haifa, Israel  
noam.ordan@gmail.com

Shuly Wintner  
Dept. of Computer Science  
University of Haifa  
31905 Haifa, Israel  
shuly@cs.haifa.ac.il

Abstract
Translation models used for statistical machine translation are compiled from parallel corpora; such corpora are manually translated, but the direction of translation is usually unknown, and is consequently ignored. However, much research in Translation Studies indicates that the direction of translation matters, as translated language (translationese) has many unique properties. Specifically, phrase tables constructed from parallel corpora translated in the same direction as the translation task perform better than ones constructed from corpora translated in the opposite direction.

We reconfirm that this is indeed the case, but emphasize the importance of using also 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 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 (Baker, 1993, 1995, 1996): translated texts tend to be shorter, to have lower type/token ratio, and to use certain discourse markers more frequently than original texts. Incidentally, translated texts are so markedly different from original ones that automatic classification can identify them with very high accuracy (van Halteren, 2008; Baroni and Bernardini, 2006; 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 used to produce those corpora. Given the unique properties of translationese, however, it is reasonable to assume that this direction may affect the quality of the translation. Recently, Kurokawa et al. (2009) showed that this is indeed the case. They train 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, whereas for translating into English it is better to use the former.

Usually, of course, the translation direction of a parallel corpus is unknown. Therefore, Kurokawa et al. (2009) train an SVM-based classifier to predict which side of a bi-text is the origin and which one is the translation, and only use the subset of the corpus that corresponds to the translation direction of the task in training their translation model.
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 an automatic evaluation metric, BLEU (Papineni et al., 2002), as well as through a qualitative analysis of the results.

Our departure point is the results of Kurokawa et al. (2009), which we successfully replicate in Section 3. First (Section 4), we 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.

Then (Section 5), we explore ways to build a translation model that is adapted to the unique properties of translationese. We first show 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. 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 the quality of the translation.

Specifically, we use the entire corpus, create a single, unified phrase table and then use the statistical measures mentioned above, 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 yield the best results, 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.

## 2 Related Work

Kurokawa et al. (2009) are the first to address the direction of translation in the context of SMT. Their main finding is that using the \( S \rightarrow T \) portion of the parallel corpus results in much better translation quality than when the \( T \rightarrow S \) portion is used for training the translation model. We indeed replicate these results here (Section 3), and view them as a baseline. Additionally, we show that the \( T \rightarrow S \) portion is also important for machine translation and thus should not be discarded. Using information-theory measures, and in particular cross-entropy, we gain statistically significant improvements in translation quality beyond the results of Kurokawa et al. (2009). Furthermore, we eliminate the need to (manually or automatically) detect the direction of translation of the parallel corpus.

Lembersky et al. (2011) also investigate the relations between translationese and machine translation. Focusing on the language model (LM), they show that LMs trained on translated texts yield better translation quality than LMs compiled from original texts. They also show that perplexity is a good discriminator between original and translated texts.

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, while Foster
et al. (2010) and Axelrod et al. (2011) use it on the translation model. 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, while we focus on an adaptation of the phrase table used for SMT. In this sense, our work follows Foster et al. (2010), who weigh out-of-domain phrase pairs according to their relevance to the target domain. They use multiple features that help 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 et al. (2009). Additionally, we apply the method suggested by Moore and Lewis (2010) using perplexity ratio instead of cross-entropy difference.

3 Experimental Setup

The tasks we focus on are translation between French and English, in both directions. We use the Hansard corpus, containing transcripts of the Canadian parliament from 1996–2007, as the source of all 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. Both English and French are lower-cased and tokenized using MOSES (Koehn et al., 2007). Sentences longer than 80 words are discarded.

To address the effect of the 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 devtest 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). French-to-English MT systems are tuned and tested on French-original sentences and English-to-French systems on English-original ones.

To replicate the results of Kurokawa et al. (2009) and set up a baseline, we train twelve French-to-English and twelve English-to-French phrase-based (PB-) SMT systems using the MOSES toolkit (Koehn et al., 2007), each trained on a different subset of the corpus. 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 construct English and French 5-gram language models from the English and French subsections of the Europarl-V6 corpus (Koehn, 2005), using interpolated modified Kneser-Ney discounting (Chen, 1998) and no cut-off on all n-grams. Europarl consists of a large number of subsets translated from various languages, and is therefore unlikely to be biased towards a specific source language. The reordering model used in all MT systems is trained on the union of the 1.5M French-original and the 1.5M English-original subsets, using msd-bidirectional-reordering. 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 ‘English-original’ (EO) and ‘French-original’ (FO) to refer to the subsets of the corpus that are translated from English to French and from French to English, respectively. The translation tasks are English-to-French (E2F) and French-to-English (F2E). We thus use ‘S → T’ when the FO corpus is used for the F2E task or when the EO corpus is used for the E2F task; and ‘T → S’ when the FO corpus is used for the E2F task or when the EO corpus is used for the F2E task.

Table 1 depicts the BLEU scores of the baseline systems. The data are consistent with the findings of Kurokawa et al. (2009): systems trained on S → T parallel texts outperform systems trained on T → S texts, even when the latter are much larger. The difference in BLEU score can be as high as 3 points.

4 Analysis of the Phrase Tables

The baseline results suggest that S → T and T → S phrase tables differ substantially, presumably due to the different characteristics of original
Table 1: BLEU scores of baseline systems

<table>
<thead>
<tr>
<th>Task: French-to-English</th>
<th>Corpus subset</th>
<th>$S \rightarrow T$</th>
<th>$T \rightarrow S$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250K</td>
<td>34.35</td>
<td>31.33</td>
</tr>
<tr>
<td></td>
<td>500K</td>
<td>35.21</td>
<td>32.38</td>
</tr>
<tr>
<td></td>
<td>750K</td>
<td>36.12</td>
<td>32.90</td>
</tr>
<tr>
<td></td>
<td>1M</td>
<td>35.73</td>
<td>33.07</td>
</tr>
<tr>
<td></td>
<td>1.25M</td>
<td>36.24</td>
<td>33.23</td>
</tr>
<tr>
<td></td>
<td>1.5M</td>
<td>36.43</td>
<td>33.73</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task: English-to-French</th>
<th>Corpus subset</th>
<th>$S \rightarrow T$</th>
<th>$T \rightarrow S$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>250K</td>
<td>27.74</td>
<td>26.58</td>
</tr>
<tr>
<td></td>
<td>500K</td>
<td>29.15</td>
<td>27.19</td>
</tr>
<tr>
<td></td>
<td>750K</td>
<td>29.43</td>
<td>27.63</td>
</tr>
<tr>
<td></td>
<td>1M</td>
<td>29.94</td>
<td>27.88</td>
</tr>
<tr>
<td></td>
<td>1.25M</td>
<td>30.63</td>
<td>27.84</td>
</tr>
<tr>
<td></td>
<td>1.5M</td>
<td>29.89</td>
<td>27.83</td>
</tr>
</tbody>
</table>

Given a source phrase $s$ and a phrase table $T$ with translations $t$ of $s$ whose probabilities are $p(t|s)$, the entropy $H$ of $s$ is:

$$H(s) = - \sum_{t \in T} p(t|s) \times \log_2 p(t|s)$$  \hspace{1cm} (1)$$

There are two major flavors of the phrase table entropy metric: Lambert et al. (2011) calculate the average entropy over all translation options for each source phrase (henceforth, phrase table entropy or $PtEnt$), whereas Koehn et al. (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 (henceforth, covering set entropy or $CovEnt$).

We also propose a metric that assesses the quality of the source side of a phrase table. The 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 (henceforth, covering set average length or $CovLen$).

Lembersky et al. (2011) show that perplexity distinguishes well between translated and original texts. Moreover, perplexity reflects the degree of ‘relatedness’ of a given phrase to original language or to translationese. Motivated by this observation, we design two cross-entropy-based measures to assess how well each phrase table fits the genre of translationese. Since MT systems are evaluated against human translations, we believe that this factor may have a significant impact on translation performance. The cross-entropy of a text $T = w_1, w_2, \cdots, w_N$ according to a language model $L$ is:

$$H(T, L) = - \frac{1}{N} \sum_{i=1}^{N} \log_2 L(w_i)$$  \hspace{1cm} (2)$$

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 out-of-vocabulary 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 ($\lambda = 0.58$). For French translationese, we use 270,000 sentences from Europarl and 3,000 sentences from Hansard, $\lambda = 0.81$. Finally, we compute the cross-entropy of each target phrase in the phrase tables according to these language models.
As with the entropy-based measures, we define two cross-entropy metrics: phrase table cross-entropy or PtCrEnt calculates the average cross-entropy over weighted cross-entropies of all translation options for each source phrase, and covering set cross-entropy or 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 2 depicts various statistical measures computed on the phrase tables corresponding to our 24 SMT systems.\(^1\) The data meet our preliminary expectations: \( S \rightarrow T \) phrase tables have more unique source phrases, but fewer translation options per source phrase. They have lower entropy and cross-entropy, but higher covering set length.

In order to assess the correspondence of each measure to translation quality, we compute the correlation of BLEU scores from Table 1 with each of the measures specified in Table 2; we compute the correlation coefficient \( R^2 \) (the square of Pearson’s product-moment correlation coefficient) by fitting a simple linear regression model. Table 3 lists the results. Only the covering set cross-entropy measure shows stability over the French-to-English and English-to-French translation tasks, with \( R^2 \) equals to 0.56 and 0.54, respectively. Other measures are sensitive to the translation task: covering set entropy has the highest correlation with BLEU (\( R^2 = 0.94 \)) when translating French-to-English, but it drops to 0.46 for the reverse task. The covering set average length measure shows similar behavior: \( R^2 \) drops from 0.75 in French-to-English to 0.56 in English-to-French. Still, the correlation of these measures with BLEU is high.

Consequently, we use the three best measures, namely covering set entropy, cross-entropy and average length, 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 Translation Model Adaptation

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 the above findings. Our goal is to establish the best method to address the issue of different translation direction components in the parallel corpus.

First, we simply take the union of the two subsets of the parallel corpus. We create three different mixes of FO and EO: 500K sentences each of FO and EO (‘MIX1’), 500K sentences of FO and 1M sentences of EO (‘MIX2’), and 1M sentences of FO and 500K sentences of EO (‘MIX3’). 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. We use the same Moses configuration as well as the same language and re-ordering models as in Section 3.

Table 4 reports the results, comparing them to the results obtained for the baseline MT systems trained on individual French-original and English-original bi-texts (see Section 3).\(^2\) Note that the mixed corpus includes many more sentences than each of the baseline models; this is a

---

\(^1\)The phrase tables were pruned, retaining only phrases that are included in the evaluation set.

\(^2\)Recall that when translating from French to English, \( S \rightarrow T \) means that the bi-text is French-original; when translating from English to French, \( S \rightarrow T \) means it is English-original.
Table 2: Statistic measures computed on the phrase tables: total size, in tokens (‘Total’); the number of unique source phrases (‘Source’); the average number of translations per source phrase (‘AvgTran’); phrase table entropy (‘PtEnt’) and covering set entropy (‘CovEnt’); phrase table cross-entropy (‘PtCrEnt’) and covering set cross-entropy (‘CovCrEnt’); and the covering set average length (‘CovLen’).}

realistic scenario, in which one can opt either to use the entire parallel corpus, or only its $S \rightarrow T$ subset. Even with a corpus several times as large, however, the ‘mixed’ MT systems perform only slightly better than the $S \rightarrow T$ ones. On one hand, this means that one can train MT systems on $S \rightarrow T$ data only, at the expense of only a minor loss in quality. On the other hand, it is obvious that the $T \rightarrow S$ component also contributes to translation quality. We now look at ways to better utilize this portion.

We compute the measures established in the previous section on phrase tables trained on the MIX corpora, and compare them with the same measures computed for phrase tables trained on the relevant $S \rightarrow T$ corpus for both translation tasks. Table 5 displays the figures for the MIX1 corpus: Phrase tables trained on mixed corpora have higher covering set average length, similar covering set entropy, but significantly worse covering set cross-entropy. Consequently, improving covering set cross-entropy has the greatest potential for improving translation quality. We therefore use this feature to ‘encourage’ the decoder to
select translation options that are more related to the genre of translated texts.

\[
F(t) = \frac{H(t, L_o)}{H(t, L_t)} \tag{4}
\]

We apply these techniques to the French-to-English and English-to-French phrase tables built from the mixed corpora and use each phrase table to train an SMT system. Table 6 summarizes the performance of these systems. All systems outperform the corresponding Union systems. ‘CrEnt’ systems show significant improvements \((p < 0.05)\) on balanced scenarios (‘MIX1’) and on scenarios biased towards the \(S \rightarrow T\) component (‘MIX2’ in the French-to-English task, ‘MIX3’ in English-to-French). ‘PplRatio’ systems exhibit more consistent behavior, showing small, but statistically significant improvement \((p < 0.05)\) in all scenarios.

Table 4: Evaluation of the MIX systems

<table>
<thead>
<tr>
<th>Task: French-to-English</th>
<th>System</th>
<th>MIX1</th>
<th>MIX2</th>
<th>MIX3</th>
<th>Union</th>
<th>(S \rightarrow T)</th>
<th>(T \rightarrow S)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Union</td>
<td>35.27</td>
<td>35.36</td>
<td>35.94</td>
<td></td>
<td>35.21</td>
<td>32.38</td>
</tr>
<tr>
<td></td>
<td>(S \rightarrow T)</td>
<td>35.21</td>
<td>35.21</td>
<td>35.73</td>
<td></td>
<td>35.21</td>
<td>32.38</td>
</tr>
<tr>
<td></td>
<td>(T \rightarrow S)</td>
<td>32.38</td>
<td>33.07</td>
<td>32.38</td>
<td></td>
<td>33.07</td>
<td>32.38</td>
</tr>
</tbody>
</table>

Table 5: Statistical measures computed for mixed vs. source-to-target phrase tables

<table>
<thead>
<tr>
<th>Measure</th>
<th>French-to-English</th>
<th>MIX1</th>
<th>(S \rightarrow T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CovLen</td>
<td>2.78</td>
<td>2.64</td>
<td></td>
</tr>
<tr>
<td>CovEnt</td>
<td>0.37</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>CovCrEnt</td>
<td>1.58</td>
<td>1.10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measure</th>
<th>English-to-French</th>
<th>MIX1</th>
<th>(S \rightarrow T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CovLen</td>
<td>2.40</td>
<td>2.25</td>
<td></td>
</tr>
<tr>
<td>CovEnt</td>
<td>0.55</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>CovCrEnt</td>
<td>2.09</td>
<td>1.48</td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Evaluation of MT Systems

<table>
<thead>
<tr>
<th>Task: French-to-English</th>
<th>System</th>
<th>MIX1</th>
<th>MIX2</th>
<th>MIX3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union</td>
<td>35.27</td>
<td>35.36</td>
<td>35.94</td>
<td></td>
</tr>
<tr>
<td>CrEnt</td>
<td>35.54</td>
<td>35.45</td>
<td>36.75</td>
<td></td>
</tr>
<tr>
<td>PplRatio</td>
<td>35.59</td>
<td>35.78</td>
<td>36.22</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task: English-to-French</th>
<th>System</th>
<th>MIX1</th>
<th>MIX2</th>
<th>MIX3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Union</td>
<td>29.27</td>
<td>30.01</td>
<td>29.44</td>
<td></td>
</tr>
<tr>
<td>CrEnt</td>
<td>29.47</td>
<td>30.44</td>
<td>29.45</td>
<td></td>
</tr>
<tr>
<td>PplRatio</td>
<td>29.65</td>
<td>30.34</td>
<td>29.62</td>
<td></td>
</tr>
</tbody>
</table>

Note again that all systems in the same column 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 op-
tions that are closer to translated texts than to original ones.

6 Analysis

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 ‘PplRatio’ (henceforth adapted) French-English SMT systems. We manually inspect 200 sentences of length between 15 and 25 from the French-English evaluation set.

In many cases, the adapted system produces more fluent and accurate translations. In the following examples, the baseline system generates common translations of French words that are adequate for a wider context, whereas the adapted system chooses less common, but more suitable translations:

**Source** J'ai eu cette perception et j'étais assez certain que ça allait se faire.
**Baseline** I had that perception and I was enough certain it was going to do.
**Adapted** I had that perception and I was quite certain it was going to do.

**Source** J'attends donc que vous en demandiez la permission, monsieur le Président.
**Baseline** I await, then, that you seek permission, mr. chairman.
**Adapted** I await, then, that you seek permission, mr. chairman.

In quite a few cases, the baseline system leaves out important words from the source sentence, producing ungrammatical, even illegible translations, whereas the adapted system generates good translations. Careful traceback reveals that the baseline system ‘splits’ the source sentence into phrases differently (and less optimally) than the adapted system. Apparently, when the decoder is coerced to select translation options that are more adapted to translationese, it tends to select source phrases that are more related to original texts, resulting in more successful coverage of the source sentence:

**Source** Pourtant, lorsqu'on les avait présentés, c'était pour corriger les problèmes liés au PCSRA.
**Baseline** Yet when they had presented, it was to correct the problems the CAIS program.
**Adapted** Yet when they had presented, it was to correct the problems associated with CAIS.

Finally, there are often cultural differences between languages, specifically the use of a 24-hour clock (common in French) vs. a 12-hour clock (common in English). The adapted system is more consistent in translating the former to the latter:

**Source** On avait décidé de poursuivre la séance jusqu'à 18 heures, mais on n'aura pas le temps de faire un autre tour de table.
**Baseline** We had decided to continue the meeting until 18 hours, but we will not have the time to do another round.
**Adapted** We had decided to continue the meeting until 6 p.m., but we won't have the time to do another round.

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 paper reflect only partly the human situation, but they do so in a crucial way. The S → 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 → S systems are flipped directions of humans’ input and output. The S → T direction proved to be more fluent, accurate and even more culturally sensitive. This has to do with fact that the translators ‘cover’ the source texts more fully, having a better ‘translation model’.

**Source** Cependant, je pense qu’il est prématuré de le faire actuellement, étant donné que le ministre a lancé cette tournée.
**Baseline** However, I think it is premature to the right now, since the minister launched this tour.
**Adapted** However, I think it is premature to do so now, given that the minister has launched this tour.

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 paper reflect only partly the human situation, but they do so in a crucial way. The S → 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 → S systems are flipped directions of humans’ input and output. The S → T direction proved to be more fluent, accurate and even more culturally sensitive. This has to do with fact that the translators ‘cover’ the source texts more fully, having a better ‘translation model’.
7 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 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.

This work can be extended in various directions. We plan to further explore the use of two phrase tables, one for each direction-determined subset of the parallel corpus. Specifically, we will interpolate the translation models as in Foster and Kuhn (2007), including a maximum a posteriori combination (Bacchiani et al., 2006). We also plan to upweight the $S \rightarrow T$ subset of the parallel corpus and train a single phrase table on the concatenated corpus. Finally, we intend to extend this work by combining the translation-model adaptation we present here with the language-model adaptation suggested by Lembersky et al. (2011) in a unified system that is more tuned to generating translationese.

Acknowledgments

We are grateful to Cyril Goutte, George Foster and Pierre Isabelle for providing us with an annotated version of the Hansard corpus. 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


George Foster, Cyril Goutte, and Roland Kuhn. Discriminative instance weighting for domain adaptation in statistical machine translation. In


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