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.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 255–265,
Avignon, France, April 23 - 27 2012.

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 mucqqh 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 lowercased 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-fe* 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 '$S \rightarrow T$' when the FO corpus is used for the F2E task or when the EO corpus is used for the E2F task; and '$T \rightarrow S$' 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 $S \rightarrow T$ parallel texts outperform systems trained on $T \rightarrow S$ 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 $S \rightarrow T$ and $T \rightarrow S$ phrase tables differ substantially, presumably due to the different characteristics of original

Task: French-to-English		
Corpus subset	$S \rightarrow T$	$T \rightarrow S$
250K	34.35	31.33
500K	35.21	32.38
750K	36.12	32.90
1M	35.73	33.07
1.25M	36.24	33.23
1.5M	36.43	33.73
Task: English-to-French		
Corpus subset	$S \rightarrow T$	$T \rightarrow S$
250K	27.74	26.58
500K	29.15	27.19
750K	29.43	27.63
1M	29.94	27.88
1.25M	30.63	27.84
1.5M	29.89	27.83

Table 1: BLEU scores of baseline systems

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, while a small number of translations per source phrase means a lower phrase table entropy. Entropy-based measures are well-established tools to assess the quality of a phrase table. Phrase table entropy captures the amount of uncertainty involved in choosing candidate translation phrases (Koehn et al., 2009).

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:

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

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:

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

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:

$$W(s, L) = -\sum_{t \in T} H(t, L) \times p(t|s) \qquad (3)$$

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 asses 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.

[1]The phrase tables were pruned, retaining only phrases that are included in the evaluation set.

Measure	R^2 (FR–EN)	R^2 (EN-FR)
AvgTran	0.06	0.22
PtEnt	0.03	0.19
CovEnt	**0.94**	0.46
PtCrEnt	0.33	0.44
CovCrEnt	0.56	0.54
CovLen	0.75	**0.56**

Table 3: Correlation of BLEU scores with phrase table statistical measures

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 mixtures 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 reordering 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

[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.

Task: French-to-English								
Set	Total	Source	AvgTran	PtEnt	CovEnt	PtCrEnt	CovCrEnt	CovLen
$S \rightarrow T$								
250K	231K	69K	3.35	0.86	0.36	3.94	1.64	2.44
500K	360K	86K	4.21	0.98	0.35	3.52	1.30	2.64
750K	461K	96K	4.81	1.05	0.35	3.24	1.10	2.77
1M	544K	103K	5.27	1.10	0.34	3.09	0.99	2.85
1.25M	619K	109K	5.66	1.14	0.34	2.98	0.91	2.92
1.5M	684K	114K	6.01	1.18	0.33	2.90	0.85	2.97
$T \rightarrow S$								
250K	199K	55K	3.65	0.92	0.45	4.00	1.87	2.25
500K	317K	69K	4.56	1.05	0.43	3.57	1.52	2.42
750K	405K	78K	5.19	1.12	0.43	3.39	1.35	2.53
1M	479K	85K	5.66	1.16	0.42	3.21	1.21	2.61
1.25M	545K	90K	6.07	1.20	0.41	3.11	1.12	2.67
1.5M	602K	94K	6.43	1.24	0.41	3.04	1.07	2.71
Task: English-to-French								
Set	Total	Source	AvgTran	PtEnt	CovEnt	PtCrEnt	CovCrEnt	CovLen
$S \rightarrow T$								
250K	224K	49K	4.52	1.07	0.63	3.48	1.88	2.08
500K	346K	61K	5.64	1.21	0.59	3.08	1.49	2.25
750K	437K	68K	6.39	1.29	0.57	2.91	1.33	2.33
1M	513K	74K	6.95	1.34	0.55	2.75	1.18	2.41
1.25M	579K	78K	7.42	1.38	0.54	2.63	1.09	2.46
1.5M	635K	81K	7.83	1.41	0.53	2.58	1.03	2.50
$T \rightarrow S$								
250K	220K	46K	4.75	1.12	0.63	3.62	2.09	2.02
500K	334K	57K	5.82	1.24	0.60	3.24	1.70	2.16
750K	421K	64K	6.54	1.31	0.58	2.97	1.48	2.25
1M	489K	69K	7.10	1.36	0.57	2.84	1.35	2.32
1.25M	550K	73K	7.56	1.40	0.55	2.74	1.25	2.37
1.5M	603K	76K	7.92	1.43	0.55	2.66	1.17	2.41

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

Task: French-to-English			
System	**MIX1**	**MIX2**	**MIX3**
Union	**35.27**	**35.36**	**35.94**
$S \rightarrow T$	35.21	35.21	35.73
$T \rightarrow S$	32.38	33.07	32.38
Task: English-to-French			
System	**MIX1**	**MIX2**	**MIX3**
Union	**29.27**	**30.01**	**29.44**
$S \rightarrow T$	29.15	29.94	29.15
$T \rightarrow S$	27.19	27.19	27.88

Table 4: Evaluation of the MIX systems

select translation options that are more related to the genre of translated texts.

French-to-English		
Measure	MIX1	$S \rightarrow T$
CovLen	2.78	2.64
CovEnt	0.37	0.35
CovCrEnt	1.58	1.10
English-to-French		
Measure	MIX1	$S \rightarrow T$
CovLen	2.40	2.25
CovEnt	0.55	0.58
CovCrEnt	2.09	1.48

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

We do so by adding to each phrase pair in the phrase tables an additional factor, as a measure of its fitness to the genre of translationese. 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*). Since cross-entropy is '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:

$$F(t) = \frac{H(t, L_o)}{H(t, L_t)} \quad (4)$$

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.

Task: French-to-English			
System	**MIX1**	**MIX2**	**MIX3**
Union	35.27	35.36	35.94
CrEnt	35.54	35.45	**36.75**
PplRatio	**35.59**	**35.78**	36.22
Task: English-to-French			
System	**MIX1**	**MIX2**	**MIX3**
Union	29.27	30.01	29.44
CrEnt	29.47	**30.44**	29.45
PplRatio	**29.65**	30.34	**29.62**

Table 6: Evaluation of MT Systems

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 do.*

Adapted *I had that perception and I was* ***quite*** *certain it was going do.*

Source ***J'attends*** *donc que vous en demandiez la permission, monsieur le Président.*

Baseline ***I look*** *so 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.*

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.*

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.*

Source *Vu qu'il est* ***17h 20****, je suis d'accord pour qu'on ne discute pas de ma motion immédiatement.*

Baseline *Seen that it is* ***17h 20****, I agree that we are not talking about my motion immediately.*

Adapted *Given that it is* ***5:20****, I agree that we are not talking about my motion immediately.*

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 $S \rightarrow T$ 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, 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

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362. Association for Computational Linguistics, July 2011. URL http://www.aclweb.org/anthology/D11-1033.

Michiel Bacchiani, Michael Riley, Brian Roark, and Richard Sproat. MAP adaptation of stochastic grammars. *Computer Speech and Language*, 20:41–68, January 2006. ISSN 0885-2308. doi: 10.1016/j.csl.2004.12.001. URL http://dl.acm.org/citation.cfm?id=1648820.1648854.

Mona Baker. Corpus linguistics and translation studies: Implications and applications. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Text and technology: in honour of John Sinclair*, pages 233–252. John Benjamins, Amsterdam, 1993.

Mona Baker. Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2):223–243, September 1995.

Mona Baker. Corpus-based translation studies: The challenges that lie ahead. In Gill Francis Mona Baker and Elena Tognini-Bonelli, editors, *Terminology, LSP and Translation. Studies in language engineering in honour of Juan C. Sager*, pages 175–186. John Benjamins, Amsterdam, 1996.

Marco Baroni and Silvia Bernardini. A new approach to the study of Translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, 21(3):259–274, September 2006. URL http://llc.oxfordjournals.org/cgi/content/short/21/3/259?rss=1.

Alison Beeby. Direction of translation (directionality). In Mona Baker and Gabriela Saldanha, editors, *Routledge Encyclopedia of Translation Studies*, pages 84–88. Routledge (Taylor and Francis), New York, 2nd edition, 2009.

Stanley F. Chen. An empirical study of smoothing techniques for language modeling. Technical report 10-98, Computer Science Group, Harvard University, November 1998.

George Foster and Roland Kuhn. Mixture-model adaptation for SMT. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 128–135. Association for Computational Linguistics, June 2007. URL http://www.aclweb.org/anthology/W/W07/W07-0717.

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

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 451–459, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1870658.1870702.

Jianfeng Gao, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 1:3–33, March 2002. ISSN 1530-0226. doi: http://doi.acm.org/10.1145/595576.595578. URL http://doi.acm.org/10.1145/595576.595578.

Martin Gellerstam. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*, pages 88–95. CWK Gleerup, Lund, 1986.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. Identification of translationese: A machine learning approach. In Alexander F. Gelbukh, editor, *Proceedings of CICLing-2010: 11th International Conference on Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 503–511. Springer, 2010. ISBN 978-3-642-12115-9. URL http://dx.doi.org/10.1007/978-3-642-12116-6.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. Improving translation quality by discarding most of the phrasetable. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 967–975. Association for Computational Linguistics, June 2007. URL http://www.aclweb.org/anthology/D/D07/D07-1103.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand, 2005. AAMT. URL http://mt-archive.info/MTS-2005-Koehn.pdf.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-2045.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 462 machine translation systems for Europe. In *Machine Translation Summit XII*, 2009.

Moshe Koppel and Noam Ordan. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1132.

David Kurokawa, Cyril Goutte, and Pierre Isabelle. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, 2009.

Patrik Lambert, Holger Schwenk, Christophe Servan, and Sadaf Abdul-Rauf. Investigations on translation model adaptation using monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 284–293. Association for Computational Linguistics, July 2011. URL http://www.aclweb.org/anthology/W11-2132.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. Language models for machine translation: Original vs. translated texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK, July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1034.

Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference, Short Papers*, pages 220–224, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1858842.1858883.

Franz Josef Och. 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, USA, 2003. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075096.1075117.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 440–447, Morristown,

NJ, USA, 2000. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1075218.1075274.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, USA, 2002. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1073083.1073135.

Nataša Pavlović. Directionality in translation and interpreting practice. Report on a questionnaire survey in Croatia. *Forum*, 5(2):79–99, 2007.

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

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

Hans van Halteren. Source language markers in EUROPARL translations. In *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*, pages 937–944, Morristown, NJ, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6.