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.

* Department of Computer Science, University of Haifa, 31905 Haifa, Israel.
E-mail: gennadi.lembersky@nice.com.
** Department of Computer Science, University of Haifa, 31905 Haifa, Israel.
E-mail: noam.ordan@gmail.com.
† Department of Computer Science, University of Haifa, 31905 Haifa, Israel.
E-mail: shuly@cs.haifa.ac.il.

Submission received: 23 June 2012; revised submission received: 13 November 2012; accepted for publication: 18 January 2013.

doi:10.1162/COLI_a_00159

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.

2011, 2012b). 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 construct 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 randomly 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 1
Europarl corpus size, in sentences and tokens.

	Original language	#Sentence	#Tokens
FR-EN	French	168,818	4,995,397
	English	134,318	3,441,120
DE-EN	German	200,037	5,571,202
	English	129,309	3,283,298
IT-EN	Italian	69,270	2,535,225
	English	125,640	3,389,736

Table 2
BLEU scores of the Europarl baseline systems.

Task	$S \rightarrow T$	$T \rightarrow S$
FR-EN	33.64	30.88
EN-FR	32.11	30.35
DE-EN	26.53	23.67
EN-DE	16.96	16.17
IT-EN	28.70	26.84
EN-IT	23.81	21.28

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 *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).

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

Table 3
BLEU scores of the Hansard baseline systems.

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

Table 4
Statistic measures computed on the phrase tables: total size, in tokens (Total), the number of unique source phrases (Source), and the average number of translations per source phrase (AvgTran).

Task: French-to-English						
Set	$S \rightarrow T$			$T \rightarrow S$		
	Total	Source	AvgTran	Total	Source	AvgTran
250K	231K	69K	3.35	199K	55K	3.65
500K	360K	86K	4.21	317K	69K	4.56
750K	461K	96K	4.81	405K	78K	5.19
1M	544K	103K	5.27	479K	85K	5.66
1.25M	619K	109K	5.66	545K	90K	6.07
1.5M	684K	114K	6.01	602K	94K	6.43

Task: English-to-French						
Set	$S \rightarrow T$			$T \rightarrow S$		
	Total	Source	AvgTran	Total	Source	AvgTran
250K	224K	49K	4.52	220K	46K	4.75
500K	346K	61K	5.64	334K	57K	5.82
750K	437K	68K	6.39	421K	64K	6.54
1M	513K	74K	6.95	489K	69K	7.10
1.25M	579K	78K	7.42	550K	73K	7.56
1.5M	635K	81K	7.83	603K	76K	7.92

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.[2] 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:

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

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 *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 *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 $w = w_1, w_2, \cdots w_N$ is:

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

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

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

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:

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

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

Table 5
Entropy-based measures computed on the phrase tables: covering set entropy (CovEnt), covering set cross-entropy (CovCrEnt), and covering set average length (CovLen).

Task: French-to-English						
Set	$S \rightarrow T$			$T \rightarrow S$		
	CovEnt	CovCrEnt	CovLen	CovEnt	CovCrEnt	CovLen
250K	0.36	1.64	2.44	0.45	1.87	2.25
500K	0.35	1.30	2.64	0.43	1.52	2.42
750K	0.35	1.10	2.77	0.43	1.35	2.53
1M	0.34	0.99	2.85	0.42	1.21	2.61
1.25M	0.34	0.91	2.92	0.41	1.12	2.67
1.5M	0.33	0.85	2.97	0.41	1.07	2.71

Task: English-to-French						
Set	$S \rightarrow T$			$T \rightarrow S$		
	CovEnt	CovCrEnt	CovLen	CovEnt	CovCrEnt	CovLen
250K	0.63	1.88	2.08	0.63	2.09	2.02
500K	0.59	1.49	2.25	0.60	1.70	2.16
750K	0.57	1.33	2.33	0.58	1.48	2.25
1M	0.55	1.18	2.41	0.57	1.35	2.32
1.25M	0.54	1.09	2.46	0.55	1.25	2.37
1.5M	0.53	1.03	2.50	0.55	1.17	2.41

Table 6
Correlation of BLEU scores with phrase table statistical measures.

Measure	R^2 (FR-EN)	R^2 (EN-FR)
CovEnt	**0.94**	0.46
CovCrEnt	0.56	0.54
CovLen	0.75	**0.56**

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 7
Evaluation results of various ways for combining phrase tables.

System	FR-EN	EN-FR	DE-EN	EN-DE	IT-EN	EN-IT
$S \rightarrow T$	33.64	32.11	26.53	16.96	28.70	23.81
UNION	33.79	32.24	26.76	17.36	29.12	23.70
MULTI-PATH	33.81	31.95	26.68	17.39	29.11	23.80
PPLMIN-1	33.86	32.47	**26.83**	**17.80**	29.23	23.86
PPLMIN-2	**33.95**	**32.65**	26.77	17.65	**29.44**	**24.01**

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 λ_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:

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

To adapt an interpolated translation model to a specific (development) corpus, let $\tilde{p}(s, t)$ 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 $\tilde{p}$ is defined as:

$$H = -\sum_{s,t} \tilde{p}(s, t) \times log_2 p(s \mid t) \tag{5}$$

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

$$\hat{\lambda} = \arg\min_{\lambda} -\sum_{s,t} \tilde{p}(s, t) \times log_2 \left(\sum_{i=1}^{n} \lambda_i p(s \mid t) \right) \tag{6}$$

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 $lex(t \mid s)$ and $lex(s \mid t)$) 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 8
MultEval scores for UNION and PPLMIN-2 systems.

Metric	System	Avg	p-value
	French-to-English		
BLEU ↑	UNION	33.7	-
	PPLMIN-2	33.9	0.0001
METEOR ↑	UNION	35.7	-
	PPLMIN-2	35.8	0.0001
TER ↓	UNION	49.7	-
	PPLMIN-2	49.5	0.0001
Length	UNION	99.4	-
	PPLMIN-2	99.5	0.0003
	English-to-French		
BLEU ↑	UNION	32.3	-
	PPLMIN-2	32.6	0.0001
METEOR ↑	UNION	53.8	-
	PPLMIN-2	54.0	0.0001
TER ↓	UNION	52.6	-
	PPLMIN-2	52.5	0.004
Length	UNION	98.7	-
	PPLMIN-2	98.9	0.0001

2008; Ilisei et al. 2010; Koppel and Ordan 2011). 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:

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

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.

Task: French-to-English			
System	**MIX**	**MIX-EO**	**MIX-FO**
$S \rightarrow T$	35.21	35.21	35.73
UNION	35.27	35.36	35.94
PPLMIN-1	35.46	35.59	36.26
PPLMIN-2	**35.75**	35.65	36.20
CrEnt	35.54	35.45	**36.75**
PplRatio	35.59	**35.78**	36.22

Task: English-to-French			
System	**MIX**	**MIX-FO**	**MIX-EO**
$S \rightarrow T$	29.15	29.15	29.94
UNION	29.27	29.44	30.01
PPLMIN-1	29.64	29.94	29.65
PPLMIN-2	29.50	**30.45**	30.12
CrEnt	29.47	29.45	**30.44**
PplRatio	**29.65**	29.62	30.34

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.

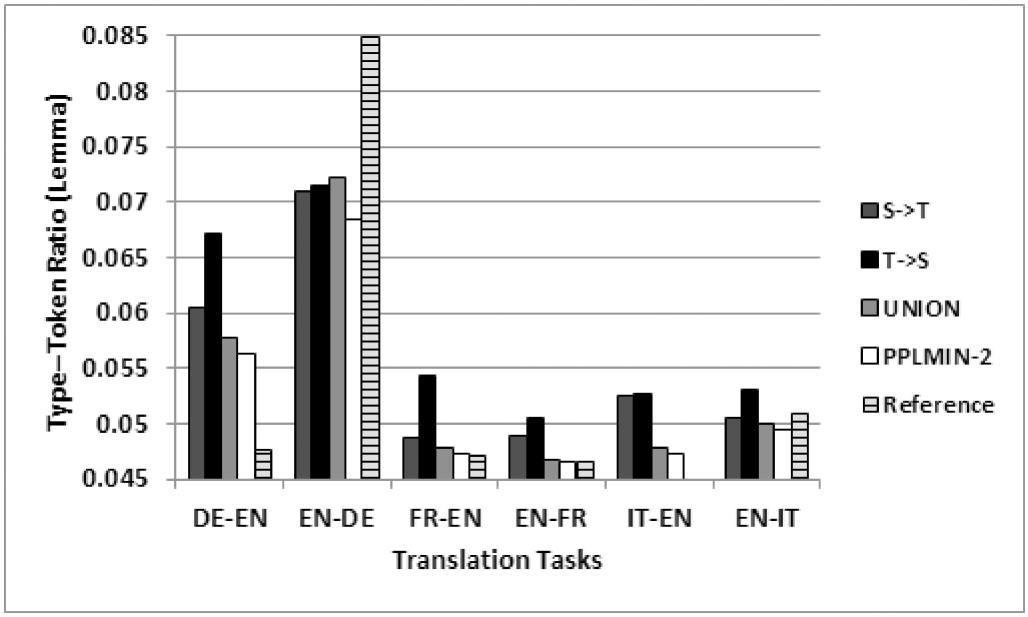

Figure 1
Type–token ratio in SMT translation outputs.

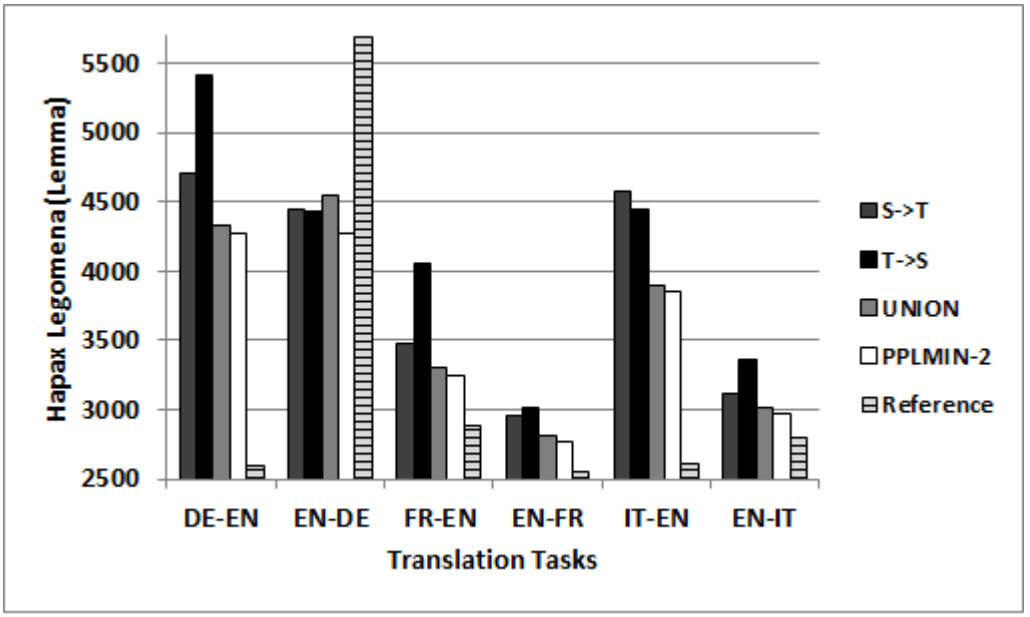

Figure 2
Numbers of singletons in SMT translation outputs.

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.

	System	CovEnt	CovCrEnt	CovLen
FR-EN	$S \rightarrow T$	0.43	2.39	2.24
	$T \rightarrow S$	0.45	2.77	2.03
	UNION	0.43	2.20	2.34
	PPLMIN-2	0.43	**2.14**	**2.35**
EN-FR	$S \rightarrow T$	0.64	3.47	2.01
	$T \rightarrow S$	0.66	3.52	1.99
	UNION	0.61	3.09	2.17
	PPLMIN-2	0.61	**2.99**	**2.18**

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:

$$MOR(S) = \frac{1}{k} \sum_{i=1}^{k} log(C(s_i)) \tag{8}$$

$C(s_i)$ is calculated from the corpus R if $s_i \in R$. Otherwise, $C(s_i) = \alpha$, where α is a predefined 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.

Task	UNION		PPLMIN-2	
	Bottom	Upper	Bottom	Upper
DE-EN	24.05	28.72	24.10	28.71
EN-DE	16.06	18.78	16.42	18.82
FR-EN	31.48	35.49	31.85	35.49
EN-FR	28.97	34.83	29.30	35.58
IT-EN	26.07	31.75	26.43	31.97
EN-IT	21.57	25.79	21.97	25.99

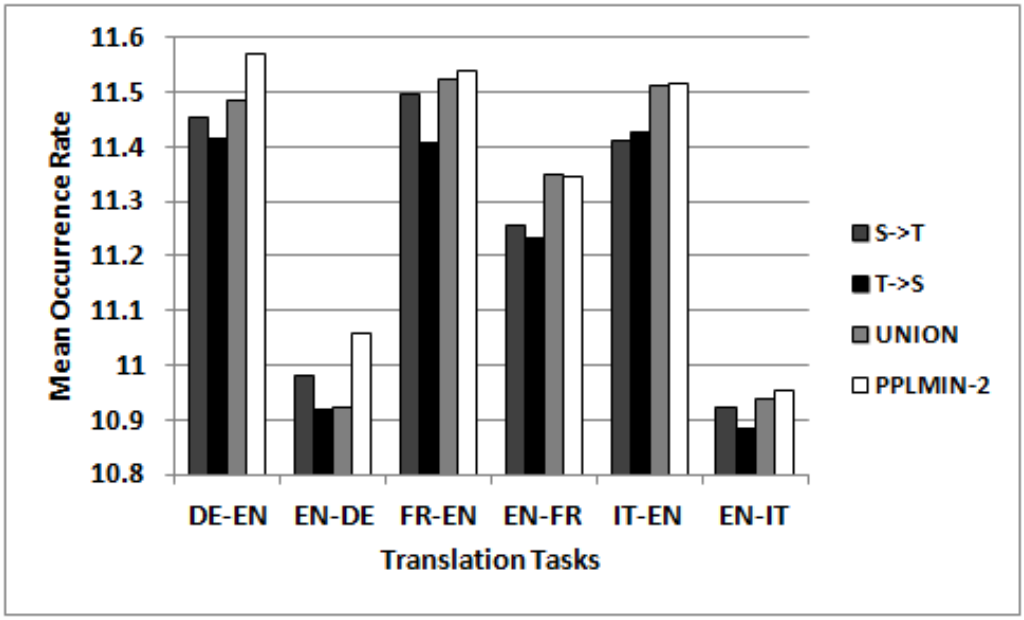

Figure 3
Mean Occurrence Rate in SMT translation outputs.

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 eur 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. The second example, on the other hand, misses the point altogether, and the third one is a clear case of interference, where the French *laisser conclure* is transferred verbatim as *leave conclusion*.

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 á 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 paient leur voiture, leurs taxes nationales, leur pot catalytique, leurs taxes sur les carburants, et* ***paient 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 fed, 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 12
Combining TMs and LMs: SMT system evaluation results.

	FR-EN (EUROPARL)				FR-EN (HANSARD)		
		LM				LM	
		O	T			O	T
TM	$T \rightarrow S$	27.06	27.30	TM	$T \rightarrow S$	24.41	25.47
	$S \rightarrow T$	30.38	30.65		$S \rightarrow T$	25.46	26.44

	EN-FR (EUROPARL)				EN-FR (HANSARD)		
		LM				LM	
		O	T			O	T
TM	$T \rightarrow S$	22.33	22.71	TM	$T \rightarrow S$	15.88	16.34
	$S \rightarrow T$	25.11	24.94		$S \rightarrow T$	17.08	17.48

Table 13
Adapting TMs and LMs: SMT system evaluation results.

	FR-EN (EUROPARL)				FR-EN (HANSARD)		
		LM				LM	
		Concat	Adapt		Concat	Adapt
TM	Concat	30.76	30.69	TM	Concat	27.65	27.48
	Adapt	31.06	**31.13**		Adapt	**27.76**	27.73

	EN-FR (EUROPARL)				EN-FR (HANSARD)		
		LM				LM	
		Concat	Adapt		Concat	Adapt
TM	Concat	25.55	25.51	TM	Concat	**18.69**	18.46
	Adapt	25.64	**25.69**		Adapt	18.65	18.68

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 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 concatentation 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

Al-Shabab, Omar S. 1996. *Interpretation and the Language of Translation: Creativity and Conventions in Translation*. Janus, Edinburgh.

Axelrod, Amittai, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh.

Baker, Mona. 1993. 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*. John Benjamins, Amsterdam, pages 233–252.

Baroni, Marco and Silvia Bernardini. 2006. 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.

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

Blum-Kulka, Shoshana. 1986. Shifts of cohesion and coherence in translation. In Juliane House and Shoshana Blum-Kulka, editors, *Interlingual and Intercultural Communication Discourse and Cognition in Translation and Second Language Acquisition Studies*, volume 35. Gunter Narr Verlag, pages 17–35.

Blum-Kulka, Shoshana and Eddie A. Levenston. 1983. Universals of lexical simplification. In Claus Faerch and Gabriele Kasper, editors, *Strategies in Interlanguage Communication*. Longman, pages 119–139.

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

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer

instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, OR.
Denkowski, Michael and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91, Edinburgh.
Ferraresi, Adriano, Silvia Bernardini, Picci Giovanni, and Marco Baroni. 2008. Web corpora for bilingual lexicography, a pilot study of English/French collocation extraction and translation. In *Proceedings of The International Symposium on Using Corpora in Contrastive and Translation Studies*, pages 1–30, Hangzhou.
Foster, George, Cyril Goutte, and Roland Kuhn. 2010. 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.
Gao, Jianfeng, Joshua Goodman, Mingjing Li, and Kai-Fu Lee. 2002. Toward a unified approach to statistical language modeling for Chinese. *ACM Transactions on Asian Language Information Processing*, 1:3–33.
Gellerstam, Martin. 1986. Translationese in Swedish novels translated from English. In Lars Wollin and Hans Lindquist, editors, *Translation Studies in Scandinavia*. CWK Gleerup, Lund, pages 88–95.
Ilisei, Iustina, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. 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.
Johnson, Howard, Joel Martin, George Foster, and Roland Kuhn. 2007. 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, Prague.
Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP 2004*, pages 388–395, Barcelona.
Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit*, pages 79–86, Phuket Island.
Koehn, Philipp, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for Europe. In *Proceedings of the Twelfth Machine Translation Summit*, pages 65–72, Ottawa.
Koehn, Philipp, 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. 2007. 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.
Koehn, Philipp and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 224–227, Stroudsburg, PA.
Koppel, Moshe and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, OR.
Kurokawa, David, Cyril Goutte, and Pierre Isabelle. 2009. Automatic detection of translated text and its impact on machine translation. In *Proceedings of MT-Summit XII*, pages 81–88, Ottawa.
Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English lexical prose. *Meta*, 43(4):557–570.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2011. 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.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2012a. Adapting translation models to translationese improves SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–265, Avignon.
Lembersky, Gennadi, Noam Ordan, and Shuly Wintner. 2012b. Language models

for machine translation: Original vs. translated texts. *Computational Linguistics*, 38(4):799–825.

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.

Och, Franz Josef and Hermann Ney. 2000. 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 Croatia. *Forum*, 5(2):79–99.

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.