Latent Domain Translation Models in Mix-of-Domains Haystack

Hoang Cuong and **Khalil Sima'an**
Institute for Logic, Language and Computation
University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands

Abstract

This paper addresses the problem of selecting adequate training sentence pairs from a mix-of-domains parallel corpus for a translation task represented by a small in-domain parallel corpus. We propose a novel latent domain translation model which includes domain priors, domain-dependent translation models and language models. The goal of learning is to estimate the probability of a sentence pair in mix-domain corpus to be in- or out-domain using in-domain corpus statistics as prior. We derive an EM training algorithm and provide solutions for estimating out-domain models (given only in- and mix-domain data). We report on experiments in data selection (intrinsic) and machine translation (extrinsic) on a large parallel corpus consisting of a *mix of a rather diverse set of domains*. Our results show that our latent domain invitation approach outperforms the existing baselines significantly. We also provide analysis of the merits of our approach relative to existing approaches.

Large parallel corpora are important for training statistical MT systems. Besides size, the *relevance* of a parallel training corpus to the translation task at hand can be decisive for system performance, cf. (Axelrod et al., 2011; Koehn and Haddow, 2012). In this paper we look at data selection where we have access to a large parallel data repository $\mathcal{C}_{mix}$, representing a rather varied mix of domains, and we are given a sample of in-domain parallel data $\mathcal{C}_{in}$, exemplifying a target translation task. Simply concatenating $\mathcal{C}_{in}$ with $\mathcal{C}_{mix}$ does not always deliver best performance, because including irrelevant sentences might be more harmful than beneficial, cf. (Axelrod et al., 2011). To make the best of available data, we must select sentences from $\mathcal{C}_{mix}$ for their relevance to translating sentences from $\mathcal{C}_{in}$.

Axelrod et al. (2011) and follow-up work, e.g., (Haddow and Koehn, 2012; Koehn and Haddow, 2012), select sentence pairs in $\mathcal{C}_{mix}$ using the cross-entropy difference between in- and mix-domain *language models*, both source and target sides, a modification of the Moore and Lewis method (Moore and Lewis, 2010). In the translation context, however, often a source phrase has different senses/translations in different domains, which cannot be distinguished with monolingual language models. The dependence of translation choice on domain suggests that the word alignments themselves can better be conditioned on domain information. However, in the data selection setting, corpus $\mathcal{C}_{mix}$ often does not contain useful domain markers, and $\mathcal{C}_{in}$ contains only a small sample of in-domain sentence pairs.

In this paper we present a *latent domain translation model* which weights every sentence pair $\langle \mathbf{f}, \mathbf{e} \rangle \in C_{mix}$ with a probability $P(D \mid \mathbf{f}, \mathbf{e})$ for being in-domain (D_1) or out-domain (D_0). Our model defines $P(\mathbf{e}, \mathbf{f}) = \sum_{D \in \{D_1, D_0\}} P(D) P(\mathbf{e}, \mathbf{f} \mid D)$, using a latent domain variable $D \in \{D_0, D_1\}$. Using bi-directional translation models, this leads to a domain prior $P(D)$, domain-dependent translation models $P_t(\cdot \mid \cdot, D)$ and language models $P_{lm}(\cdot \mid D)$ as in Equation 1:

$$P(\mathbf{e}, \mathbf{f} \mid D) \;=\; \frac{1}{2} \times \{P_{lm}(\mathbf{e} \mid D) P_t(\mathbf{f} \mid \mathbf{e}, D) \;+\; P_{lm}(\mathbf{f} \mid D) P_t(\mathbf{e} \mid \mathbf{f}, D)\} \tag{1}$$

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1928–1939, Dublin, Ireland, August 23-29 2014.

For efficiency we assume IBM Model I alignments $\mathbf{a}$ and translation tables $t(\cdot)$, e.g., $P_t(\mathbf{e} \mid \mathbf{f}, D) \propto \sum_{\mathbf{a}} \prod_i t(e_i \mid f_{\mathbf{a}_i}, D)$. Language models (LMs) P_{lm} are trained separately, albeit one problem not addressed by earlier work is how to train out-domain LMs given only in- and mix-domain data?

In our model, initially both the translation and LM probabilities estimated from $\mathcal{C}_{in}$ serve as *priors* for weighting sentence pairs in $\mathcal{C}_{mix}$ as being *more* relevant for in-domain translation *than not*. This initial weighting reveals *pseudo out-domain* data in $\mathcal{C}_{mix}$, which we use to train out-domain language models as well as initialize out-domain word alignment tables.[1] With these sharpened translation and language models, training commences using a version of EM (Dempster et al., 1977). Because the *potentially relevant data* in $\mathcal{C}_{mix}$ might be a superset of any in-domain data, the estimates from $\mathcal{C}_{in}$ serve merely as initial model estimates. Metaphorically, iterative EM training resembles `party invitations` on social networks (hence, the *Invitation model*): if initially in/out-domain sentence pairs (the `hosts`) `invite` some sentence pairs from $\mathcal{C}_{mix}$, in the next iteration the new `pseudo` in/out-domain sentences help invite more sentence pairs. In EM, sentence pairs receive weighted, rather than absolute, invitations from in- and out-domain models.

We present extensive experiments on a rather difficult selection task exploiting a large mix-domain corpus of 4.61M sentence pairs. Initially we conduct intrinsic evaluation on the mix-domain corpus where we also hide in-domain data and seek to retrieve it. Subsequently we conduct full MT experiments over the task. The results show that our Invitation model gives far better selections as well as translation performance than the baseline trained on the large data C_{mix}.

1 Invitation models of weighting and selection

By now training data selection from large mix-domain data is an accepted necessity, e.g., (Axelrod et al., 2011; Gascó et al., 2012; Haddow and Koehn, 2012; Banerjee et al., 2012; Irvine et al., 2013). Data selection has a different (but complementary) goal than domain adaptation, which aims at adapting an *existing out-domain system* by focusing on, e.g., translation model (Koehn and Schroeder, 2007; Foster and Kuhn, 2007; Sennrich, 2012), reordering model (Chen et al., 2013) and/or language model adaptation (Eidelman et al., 2012). Our setting is in line with data selection approaches (Moore and Lewis, 2010; Axelrod et al., 2011; Duh et al., 2013), and is somewhat related to phrase pair weighting (Matsoukas et al., 2009; Foster et al., 2010). In this paper we explicitly draw attention to the special case of a mix-domain parallel corpus consisting of a large and rather diverse set of domains.

Our model assigns to every sentence pair $\langle \mathbf{f}, \mathbf{e} \rangle \in \mathcal{C}_{mix}$ a probability as in Equation 2:

$$\begin{aligned} P(D \mid \mathbf{f}, \mathbf{e}) &= \frac{P(\mathbf{f}, \mathbf{e}, D)}{\sum_{D \in \{D_1, D_0\}} P(\mathbf{f}, \mathbf{e}, D)} \\ P(\mathbf{f}, \mathbf{e}, D) &= \frac{1}{2} \times P(D) \times \{P_{lm}(\mathbf{e} \mid D) P_t(\mathbf{f} \mid \mathbf{e}, D) + P_{lm}(\mathbf{f} \mid D) P_t(\mathbf{e} \mid \mathbf{f}, D)\} \end{aligned} \tag{2}$$

Viewed as learning two latent corpora $\mathcal{C}_1$ and $\mathcal{C}_0$, the task is to assign every $\langle \mathbf{f}, \mathbf{e} \rangle \in C_{mix}$ an expected count $P(D_x \mid \mathbf{f}, \mathbf{e})$ that it is in $\mathcal{C}_x \in \{C_0, C_1\}$. Next we discuss the model components each in turn.

The domain-dependent translation models $P_t(\cdot \mid D)$ can be viewed as modeling the probability that $\mathbf{e}$ translates as $\mathbf{f}$ in domain $D \in \{D_0, D_1\}$. Given $\mathbf{f} = f_1, f_2, \ldots, f_m$ and $\mathbf{e} = e_1, e_2, \ldots, e_l$, we assume (hidden) alignments $\mathbf{a} = a_1, a_2, \ldots, a_m$ akin to IBM Model I (Brown et al., 1993):

$$P_t(\mathbf{f}, \mathbf{a} \mid \mathbf{e}, D) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} t(f_j | e_{a_j}, D) \tag{3}$$

$$P_t(\mathbf{f} \mid \mathbf{e}, D) = \sum_{\mathbf{a}} P_t(\mathbf{f}, \mathbf{a} | \mathbf{e}, D) = \frac{\epsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j | e_i, D). \tag{4}$$

[1] Earlier work on data selection exploits the contrast between in-domain and mix-domain instead of (pseudo) out-domain language models. However, the mix-domain language models trained on a mix of rather diverse set of domains could be considered kind of wide-coverage, which makes for a rather weak contrast with the in-domain language models.

where $t(f_j|e_{a_j}, D)$ is the domain-dependent lexical probability of f_j given e_{a_j} with respect to D. One crucial aspect about model inspired by IBM-Model-I is that $P_t(\mathbf{f} \mid \mathbf{e}, D)$ can be estimated efficiently, as in Equation 4. This makes the training particularly efficient as detailed in Section 2

The in-/out-domain source and target language models are not the same as in previous work, e.g., (Axelrod et al., 2011), which employ in-/mix-domain language models. This makes explicit the difficulty in finding data to train out-domain language models, and we present a solution in Section 2.

The domain priors $P(D_1)$ and $P(D_0)$ represent the percentage of the pairs that are in-/ and out domain respectively in $\mathcal{C}_{mix}$ learned by our model. Their estimate during training might be a reasonable selection cut-off threshold. However, we found that it is not entirely clear whether these cut-off criteria might exclude other relevant/irrelevant pairs that are not exactly in-domain. We leave this extension for future work.[2]

Finally, it should be noted that the domain-dependent word alignment model, $t(f|e, D)$ is a generalization of the standard (domain-independent) word alignment model, $t(f|e)$, in which, $t(f|e, D) = \frac{t(f|e)t(D|f,e)}{\sum_f t(f|e)t(D|f,e)}$. Here, $t(D|f, e)$ can be thought of as the *latent word-relevance models*, i.e., the probability that a word pair is relevant for in- (D_1) or out-domain (D_0). Empirical results (beyond the scope of this work) show that training the latent in-domain alignment model, $t(f|e, D_1)$ often gives better translation systems than training the standard (domain-independent) alignment model, $t(f|e)$.

2 Training

With all language models trained separately, our selection model can be viewed to have two sets of domain-dependent parameters $\Theta = \{\Theta_{D_0}, \Theta_{D_1}\}$. The parameters Θ_D consist of the domain-dependent lexical parameters (e.g., $t_{\Theta_D}(f|e, D)$, $t_{\Theta_D}(e|f, D)$) and the domain prior parameter (e.g., $P_{\Theta_D}(D)$). Our training procedure seeks the parameters Θ that maximize the log-likelihood of $\mathcal{C}_{mix}$:

$$\mathcal{L} = \sum_{\mathbf{f},\mathbf{e}} \log P_\Theta(\mathbf{f}, \mathbf{e}) \quad = \quad \sum_{\mathbf{f},\mathbf{e}} \log \sum_D \sum_{\mathbf{a}} P_{\Theta_D}(\mathbf{a}, D, \mathbf{f}, \mathbf{e}) \tag{5}$$

Because of the latent variables $\mathbf{a}$ and D, there is no closed form solution and the model is fit using the EM algorithm (Dempster et al., 1977). EM can be seen to maximize $\mathcal{L}$ via block-coordinate ascent on a lower bound $\mathcal{F}(q, \Theta)$ using an auxiliary distribution over the latent variables $q(\mathbf{a}, D|\mathbf{f}, \mathbf{e})$

$$\mathcal{L} \quad \geq \quad \mathcal{F}(q, \Theta) = \sum_{\mathbf{f},\mathbf{e}} \sum_D \sum_{\mathbf{a}} q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e}) \log \frac{P_{\Theta_D}(\mathbf{a}, D, \mathbf{f}, \mathbf{e})}{q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e})} \tag{6}$$

where the inequality results from log being concave and Jensen's inequality. We rewrite the Free energy $\mathcal{F}(q, \Theta)$ (Neal and Hinton, 1999) as follows:

$$\begin{aligned}
\mathcal{F}(q, \Theta) \quad &= \quad \sum_{\mathbf{f},\mathbf{e}} \sum_D \sum_{\mathbf{a}} q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e}) \log \frac{P_{\Theta_D}(\mathbf{a}, D, \mathbf{f}, \mathbf{e})}{q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e})} \\
&= \sum_{\mathbf{f},\mathbf{e}} \sum_{D,\mathbf{a}} q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e}) \log \frac{P_{\Theta_D}(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e})}{q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e})} \; + \; \sum_{\mathbf{f},\mathbf{e}} \sum_{D,\mathbf{a}} q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e}) \log P_\Theta(\mathbf{f}, \mathbf{e}) \\
&= \sum_{\mathbf{f},\mathbf{e}} \log P_\Theta(\mathbf{f}, \mathbf{e}) - KL[q(\mathbf{a}, D \mid \mathbf{f}, \mathbf{e}) \;||\; P_{\Theta_D}(\mathbf{a}, D|\mathbf{f}, \mathbf{e})]
\end{aligned} \tag{7}$$

where $KL[\cdot||\cdot]$ is the KL-divergence. To find $q^*(\mathbf{a}, D|\mathbf{f}, \mathbf{e})$ that maximizes $\mathcal{F}(q, \Theta)$:

$$\begin{aligned}
q^*(\mathbf{a}, D|\mathbf{f}, \mathbf{e}) \quad &= \quad \operatorname*{argmax}_{q(\mathbf{a},D|\mathbf{f},\mathbf{e})} \mathcal{F}(q, \Theta) = \operatorname*{argmin}_{q(\mathbf{a},D|\mathbf{f},\mathbf{e})} KL[q(\mathbf{a}, D|\mathbf{f}, \mathbf{e})||P_{\Theta_D}(\mathbf{a}, D|\mathbf{f}, \mathbf{e})] \\
&= \quad P_{\Theta_D}(\mathbf{a}, D|\mathbf{f}, \mathbf{e}) = P_{\Theta_D}(D|\mathbf{f}, \mathbf{e}) P_{\Theta_D}(\mathbf{a}|\mathbf{f}, \mathbf{e}, D).
\end{aligned} \tag{8}$$

[2]We especially thank an anonymous reviewer who gave valuable comments related to this point.

Here

$$P_{\Theta_D}(\mathbf{a}|\mathbf{f},\mathbf{e},D) \;=\; \frac{P_{\Theta_D}(\mathbf{f},\mathbf{a}|\mathbf{e},D)}{P_{\Theta_D}(\mathbf{f}|\mathbf{e},D)} \;=\; \frac{\prod_{j=1}^{m} t(f_j|e_{a_j},D)}{\prod_{j=1}^{m}\sum_{i=0}^{l} t(f_j|e_i,D)} \tag{9}$$

The distribution $q^*(\mathbf{a},D|\mathbf{f},\mathbf{e})$ together with $q^*(D|\mathbf{f},\mathbf{e}) = \sum_{\mathbf{a}} q^*(\mathbf{a},D|\mathbf{f},\mathbf{e}) = P_{\Theta_D}(D|\mathbf{f},\mathbf{e})$ can be used to softly fill in the values of $\mathbf{a}$ and D respectively to estimate model parameters.

We now state our derived EM update formulas. We use the notation $P^{(c)}$ and $t^{(c)}$ for current iteration estimates, and $P^{(+)}$ and $t^{(+)}$ for the re-estimates. We denote the expected counts that e aligns to f in the translation $(\mathbf{f}|\mathbf{e})$ with respect to a domain D with $c(f|e;\mathbf{f},\mathbf{e},D)$. Similarly, we denote the expected count of $(\mathbf{f}|\mathbf{e})$ with respect to a domain D by $c(D;\mathbf{f},\mathbf{e})$.

E-step $\forall D \in \{D_0, D_1\}$ do

$$\begin{aligned} c(D;\mathbf{f},\mathbf{e}) &= P^{(c)}(D \mid \mathbf{f},\mathbf{e}) \\ c(f|e;\mathbf{f},\mathbf{e},D) &= P^{(c)}(D \mid \mathbf{f},\mathbf{e}) \frac{t^{(c)}(f \mid e,D)}{\sum_{i=0}^{l} t^{(c)}(f \mid e_i,D)} \sum_{j=1}^{m} \delta(f,f_j) \sum_{i=0}^{l} \delta(e,e_i) \end{aligned}$$

M-step $\forall D \in \{D_0, D_1\}$ do

$$t^{(+)}(f|e,D) = \frac{\sum_{\mathbf{f},\mathbf{e}} c(f|e;\mathbf{f},\mathbf{e},D)}{\sum_{f}\sum_{\mathbf{f},\mathbf{e}} c(f|e;\mathbf{f},\mathbf{e},D)} \qquad P^{(+)}(D) = \frac{\sum_{\mathbf{f},\mathbf{e}} c(D;\mathbf{f},\mathbf{e})}{\sum_{D}\sum_{\mathbf{f},\mathbf{e}} c(D;\mathbf{f},\mathbf{e})}$$

To re-estimate $P(D \mid \mathbf{f},\mathbf{e})$ we substitute the M-step estimates into Equations 3, 4 and 2. We initialize translation tables $t(f|e,D_1)$ and $t(e|f,D_1)$ with non-zero estimates obtained from applying IBM model I to in-domain corpus $\mathcal{C}_{in}$.[3] Before EM training starts we must train the LMs. The in-domain LMs $P_{lm}(\mathbf{e}|D_1)$ and $P_{lm}(\mathbf{f}|D_1)$ are trained on the source and target sides of $\mathcal{C}_{in}$ respectively. For the out-domain LMs $P_{lm}(\mathbf{e}|D_0)$ and $P_{lm}(\mathbf{f}|D_0)$ we need an *out-domain* data set to train them. It would also be reasonable to use the set to train the out-domain tables, $t(\cdot \mid \cdot, D_0)$. This raises an hitherto unattended question regarding how to construct such an out-domain data set.

Inspired by burn-in in sampling, initially we isolate all LMs from our model to train the translation models for a single EM iteration; we initialize the model with a translation table constructed on $\mathcal{C}_{in}$ and uniform otherwise. Using the re-estimates, we score sentence pairs in $\mathcal{C}_{mix}$ with $P(D_1|\mathbf{f},\mathbf{e})$ and select a *burn-in subset* of smallest scoring pairs as *pseudo out-domain data* which can be used to train $P_{lm}(\mathbf{e}|D_0)$ and $P_{lm}(\mathbf{f}|D_0)$. Choosing the optimal size of this subset is difficult, but in practice, we usually choose a subset that has similar size (number of words) to the given in-domain corpus. The rationale behind this choice is to avoid the risk that pseudo out-domain models would dominate the in-domain models during further training. We observe that choosing the same size for a pseudo out-domain corpus is not guaranteed to always give optimal performance, and this point deserves further study.

Finally, once the domain-dependent LMs have been trained, the domain-dependent LM probabilities stay *fixed* during EM. Crucially, it is important to scale the probabilities of the four LMs to make them comparable: we normalize the probability that a LM assigns to a sentence by the total probability this LM assigns to all sentences in $\mathcal{C}_{mix}$.

3 Experimental setting

We carry out experiments in data selection (intrinsic) as well as in machine translation (extrinsic). We build an English-Spanish mix-domain corpus consisting of a large and rather varied set of domains (a

[3]Note that in practice, we usually use only one iteration to train IBM Model I. To simplify the implementation, we ignore factor $\frac{\epsilon}{(l+1)^m}$ in the model (Equation 3), which serves a minor role. It should be also noted that we set a (small) threshold, e.g., $t(\cdot|\cdot,\cdot) = 0.0001$ for all word pairs that do not occur in the in-domain corpus to avoid over-fitting.

haystack) in a way that allows us to directly measure selection quality. Starting out from a general-domain corpus $\mathcal{C}_g$ consisting of $4.51M$ sentence pairs, collected from multiple resources including EuroParl (Koehn, 2005), Common Crawl Corpus, UN Corpus, News Commentary, TAUS Software, TAUS Hardware, and TAUS Pharmacy, and a 177K in-domain (TAUS Legal) sentence pairs.

We create $\mathcal{C}_{mix}$ by selecting an arbitrary 100K pairs of in-domain set and adding them to $\mathcal{C}_g$; the remaining 77K in-domain pairs constitute $\mathcal{C}_{in}$. We think of this as `hiding` in-domain data in $\mathcal{C}_{mix}$ so we can evaluate our ability to retrieve it; in this setting we can evaluate selection directly using pseudo-precision/recall defined as the percentage of selected in-domain pairs to the total selected or to the hidden 100K pairs respectively.

Table 1 summarizes the data and the translation task. It should be noted that a mix-domain corpus, that contains a large and rather varied set of domains, frequently contains subsets with a vocabulary that is close to the in-domain adaptation task; in this case, e.g., Europarl and TAUS Legal share big portions of their source vocabulary, whereas their translations could differ. This makes the selection task far more difficult than assumed by previous approaches as we will show next.

Task	**Corpora**	**English**	**Spanish**
	Mix-Domain Corpus (4.51M sents)	$125,339,057$	$139,655,311$
TAUS Legal	In-Domain Corpus (77K sents)	$1,555,342$	$1,733,370$
	Dev (2K sents)	$27,983$	$30,501$
	Test (2K sents)	$45,736$	$48,999$

Table 1: The data preparation - training, dev and testing corpora (size in words). Note that the dev set contains sentences of 10-25 words, while the test set contains sentences that vary substantially in length, from 5-10 words up to 45-50 words.

Our Invitation model takes 3 EM-iterations to train.[4] We then weigh sentence pairs under our model with $P(D_1 \mid \mathbf{e}, \mathbf{f})$. We test various baseline models, including the bilingual cross-entropy difference model, and the two cross-entropy difference models (on the source language and on the target language).[5] We report *pseudo*-precision/recall at the sentence-level using a range of cut-off criteria for selecting the top scoring instances in the mix-domain corpus.

We use Moses (Koehn et al., 2007) with GIZA++ (Och and Ney, 2003) and k-best batch MIRA (Cherry and Foster, 2012). Final MT systems use the same *non-adapted language models* trained on 2.2M English Europarl sentences plus 248.8K sentences from News Commentary Corpus (WMT 2013).

We report BLEU (Papineni et al., 2002), METEOR 1.4 (Denkowski and Lavie, 2011) and TER (Snover et al., 2006). Statistical significance uses 95% confidence intervals using paired bootstrap re-sampling (Press et al., 1992; Koehn, 2004). The k-best batch MIRA optimizer (Cherry and Foster, 2012) was run at least three times to optimize any SMT system to avoid instability (Clark et al., 2011).[6]

4 Results

Table 2 presents the results showing substantial improvement in selection performance compared to all the baselines. Subsequently we build SMT systems over the selected subsets. We report the translation yielded by these systems over the task in Table 2 as well. It can be easily seen that the baseline approaches that simply train on in- and mix-domain data do not work that well for a difficult selection task from a mix-domain corpus consisting of a large and rather diverse set of domains. The SMT sys-

[4]To train the LM probs, we construct interpolated 4-gram Kneser-Ney language models using BerkeleyLM (Pauls and Klein, 2011). This setting for training language models is used for all experiments in this work.

[5]The script we use to train these models is developed by Luke Orland and available at: `https://github.com/lukeorland/moore\_and\_lewis\_data\_selection`.

[6]Note that metric scores for the systems are averages over multiple runs.

Cut-off	Model	In-domain Pairs	pseudo-Precision	pseudo-Recall	BLEU	METEOR	TER
50K	CE Difference (source side)	370	0.74	0.37	20.5	28.0	62.3
	CE Difference (target side)	375	0.75	0.38	19.3	26.8	63.3
	Bilingual CE Difference	413	0.83	0.41	18.7	26.3	64.3
	Invitation	19156	38.31	19.16	36.5	36.4	47.1
100K	CE Difference (source side)	592	0.59	0.59	24.8	30.8	57.8
	CE Difference (target side)	572	0.57	0.57	22.1	29.7	60.1
	Bilingual CE Difference	649	0.65	0.65	23.1	30.0	58.9
	Invitation	30474	30.47	30.47	37.1	36.9	47.0
150K	CE Difference (source side)	753	0.50	0.75	26.4	32.0	56.2
	CE Difference (target side)	742	0.49	0.74	23.9	31.2	58.8
	Bilingual CE Difference	793	0.53	0.79	24.4	30.9	58.1
	Invitation	38424	25.62	38.42	37.1	37.0	46.7
200K	CE Difference (source side)	874	0.44	0.87	26.6	32.4	56.0
	CE Difference (target side)	888	0.44	0.88	25.8	32.1	57.2
	Bilingual CE Difference	932	0.93	0.65	25.7	32.0	57.0
	Invitation	44392	22.17	44.39	37.5	37.4	46.2
250K	CE Difference (source side)	994	0.40	0.99	27.3	32.8	55.4
	CE Difference (target side)	997	0.40	0.10	26.3	32.4	56.3
	Bilingual CE Difference	1062	0.42	1.06	26.6	32.7	55.6
	Invitation	49419	19.77	49.42	37.3	37.3	46.1
300K	CE Difference (source side)	1122	0.37	1.12	28.2	33.4	54.5
	CE Difference (target side)	1093	0.36	1.09	26.4	32.7	56.0
	Bilingual CE Difference	1169	0.39	1.17	27.8	33.3	54.9
	Invitation	53892	17.96	53.89	37.7	37.5	46.0

Table 2: Systematic comparison between selection models.

tems trained on the selection of our model perform significantly and consistently better (with p-value $= 0.0001$ for all cases) than the others trained on the selection of the baselines.

	Sentences
	Bilingual CE Difference
1	*by assisting in the placement and financing of used and end-of-lease aircraft , atr asset management has helped broaden atr 's customer base , notably in emerging markets , by providing quality reconditioned aircraft at attractive prices and has helped maintain residual values of used aircraft .* *al participar en la colocación y en la financiación de los aviones usados al final del período de arrendamiento , atr gestión de activos ha podido ampliar la base de su clientela , en particular en los países de economías emergentes , al proporcionar aparatos entregados en buen estado a precios interesantes y ha contribuido a mantener el valor residual de los aviones usados .*
2	*in contrast , recent improvements in western europe are not expected to be reversed significantly .* *en cambio no se espera que las recientes mejoras en europa occidental se inviertan significativamente .*
3	*creating xml file ...* *creando el archivo xml ...*
	Invitation Model
1	*as she has said , the harmonisation of the requirements for information to appear on the invoice will mean that traders operating within the single market will be subject to a single legislation , while until now they have had to know , comply with and apply fifteen different legislations .* *como ella ha dicho , la armonización de los requisitos de información que deben constar en la factura permitirá a los comerciantes que operen en el mercado interior sujetarse a una sola legislación , mientras que hasta ahora tenían que conocer , sujetarse y aplicar quince legislaciones diferentes .*
2	*the solicitation documents shall specify the estimated period of time following dispatch of the notice of acceptance that will be required to obtain the approval .* *en el pliego de condiciones se indicará el plazo de tiempo previsto , a partir de la expedición del aviso de aceptación , que será requerido para obtener la aprobación .*
3	*there is no doubt that disadvantages will result for the consumer and for the manufacturer of branded goods , for example with regard to consumer health protection .* *ello generará , sin duda alguna , desventajas para el consumidor y el productor de artículos de marca , entre otros aspectos también en lo que se refiere a la protección de la salud del consumidor .*

Table 3: Top pairs from mix-domain corpus with highest scores according to models.

Table 3 presents some random top ranked sentence pairs from the bilingual cross-entropy difference

Model	Cut-off: 50K English	Cut-off: 50K Spanish	Cut-off: 100K English	Cut-off: 100K Spanish	Cut-off: 200K English	Cut-off: 200K Spanish
CE Difference (source side)	8.65	8.70	11.92	12.21	15.50	16.22
CE Difference (target side)	8.14	10.09	11.61	14.13	15.45	18.50
Bilingual CE Difference	7.03	8.16	10.38	11.96	14.34	16.43
Invitation	40.16	44.70	37.30	41.59	34.32	38.32

Table 4: Average words in selected sentences.

model against our Invitation model for the task. This shows clearly more relevant pairs for our selection model than for the baselines. It should be noted that the baseline models tend to prefer shorter sentences, while our model suffers less from this kind of bias. Table 4 presents the average length (in words) of selected sentences selected by different models over various cut-offs.

Cut-off	Model	In-domain Pairs	pseudo-Precision	pseudo-Recall	BLEU	METEOR	TER
300K	**Without Translation Model**	34156	11.39	34.16	35.8	36.6	47.3
	Without Language Model	51991	17.33	51.99	37.4	37.4	46.6
	Full model	53892	17.96	53.89	37.7	37.5	46.0

Table 5: Experiments exploring the roles of individual components in our model.

Which component type (language or translation models) contributes more to performance? We neutralize each component in turn and build a selection system with the remaining model parameters. Table 5 shows translation models are crucial for performance, while domain-dependent LMs make a small, yet noteworthy contribution. It should also be noted that using the LMs derived separately from in- and out-domain data yields far better performance than the LMs derived from in- and mix-domain data for this task.

System	Phrases	BLEU	METEOR	TER
Large data C_{mix}	236.74M	36.8	37.2	47.1
Subset of 300K	22.47M	37.7	37.5	46.0

Table 6: Translation accuracy comparison.

Finally, we compare a system trained on a selection of the top scored 300K sentences to a baseline large-scale SMT system trained on C_{mix} (4.61M sentences). The baseline trained on C_{mix} works with 236.74M phrase pairs, whereas the Invitation trained system employs a small table of 22.47M phrases. Tabel 6 shows the results. It is interesting that the small MT system trained by Invitation performs significantly better (with p-value $= 0.0001$ for all metrics) than the large-scale system baseline trained on all of C_{mix}.

Input	*cada estado miembro **supervisará** la categoría científica de la **evaluación** y las actividades de los miembros de los comités y de los expertos que haya designado, pero se **abstendrá** de darles instrucciones incompatibles con las funciones que les competen.*
Reference	*each member state shall **monitor** the scientific level of the **evaluation** carried out and supervise the activities of members of the committees and the experts it nominates, but shall **refrain** from giving them any instruction which is incompatible with the tasks incumbent upon them.*
Large C_{mix}	*each member state will **oversee** the category scientific **assessment** and the activities of members of the committees and experts which designated, but **abstain** of instruct incompatible with their regulatory functions.*
Subset 300K	*each member state will **monitor** the scientific category of the **evaluation** and the activities of the members of the committees and of experts who has designated, but **refrain** from giving them instructions incompatible with the required functions assumed.*

Table 7: Translation example yielded by systems.

To give a sense of the improvement in translation, we present an example in Table 7. The example is indeed illuminating because it shows the difference in choice between the mix-domain system and

our selection-trained system. The example shows different translation pairs: ⟨supervisará-*monitor*⟩ vs. ⟨supervisará-*oversee*⟩, ⟨evaluación-*evaluation*⟩ vs. ⟨evaluación-*assessment*⟩, and ⟨abstendrá de-*refrain from*⟩ vs. ⟨abstendrá de-*abstain*⟩. Table 8 presents phrase table entries, i.e., $p(e \mid f)$ and $p(f \mid e)$, for the pairs of words in each system.

		supervisará		**evaluación**		**abstendrá de**	
System	**Entry**	**monitor**	**oversee**	**evaluation**	**assessment**	**refrain from**	**abstain**
Large data C_{mix}	$\phi(e\|f)$	0.002	0.020	0.579	0.429	0.002	0.013
	$\phi(f\|e)$	0.119	0.081	0.391	0.403	0.014	0.060
Subset of 300**K**	$\phi(e\|f)$	0.012	0.024	0.487	0.357	0.015	–
	$\phi(f\|e)$	0.203	0.072	0.338	0.417	0.143	–

Table 8: Phrase entry examples. Note that the system trained on the subset of top 300K pairs of sentences does not contain the phrase pair ⟨*refrain from-abstain*⟩.

5 Final Machine Translation experiments: Putting all data together

For final adaptation evaluations we follow (Koehn and Schroeder, 2007; Nakov, 2008) and (Axelrod et al., 2011; Sennrich, 2012), by passing multiple phrase tables directly to the Moses decoder and tuning a system using these different tables together. Table 9 presents the result, showing the consistent improvement of adaptation with Invitation model compared to the baselines (with p-value $= 0.0001$ for all cases) over the mixture data C_{mix}.

Data	**System**	**BLEU**	**METEOR**	**TER**
	In-domain	36.66	37.19	44.76
50K	+ CE Difference (source side)	37.1	36.7	48.1
	+ CE Difference (target side)	37.1	36.6	48.2
	+ Bilingual CE Difference	37.1	36.6	48.2
	+ **Invitation**	38.0	37.2	47.3
100K	+ CE Difference (source side)	37.3	36.8	47.9
	+ CE Difference (target side)	37.2	36.8	48.0
	+ Bilingual CE Difference	37.2	36.8	48.0
	+ **Invitation**	38.4	37.4	46.9
150K	+ CE Difference (source side)	37.1	36.9	48.2
	+ CE Difference (target side)	37.3	36.9	47.9
	+ Bilingual CE Difference	37.0	36.8	48.1
	+ **Invitation**	38.6	37.5	46.6
200K	+ CE Difference (source side)	37.3	36.9	47.7
	+ CE Difference (target side)	37.3	36.9	47.9
	+ Bilingual CE Difference	37.3	36.9	47.8
	+ **Invitation**	38.4	37.6	46.7
250K	+ CE Difference (source side)	37.4	36.9	47.7
	+ CE Difference (target side)	37.3	37.0	47.7
	+ Bilingual CE Difference	37.3	37.0	47.8
	+ **Invitation**	38.6	37.7	46.5
300K	+ CE Difference (source side)	37.3	37.0	47.8
	+ CE Difference (target side)	37.1	37.0	48.0
	+ Bilingual CE Difference	37.3	36.9	47.8
	+**Invitation**	38.9	37.9	46.3

Table 9: Translation results from our domain-adapted SMT systems.

Finally, we also test the adaptation evaluations between the system trained on the small selection of top 300K sentences against the large-scale SMT system trained on C_{mix} when combined with the in-domain trained system. Table 10 presents the results, revealing comparable translation performance, although they are trained on data sets that are significantly different in size.

System	BLEU	METEOR	TER
In-domain + Large data C_{mix}	39.0	38.0	46.3
In-domain + Subset of 300**K**	38.9	37.9	46.3

Table 10: Translation results from our domain-adapted SMT system and the large-scale SMT system. Note that the baseline is slightly better than our domain-adapted SMT system under BLEU and METEOR, however, not statistically significant.

6 Final notes on mix-domain data selection

The specific data selection scenario studied in this paper brings up different aspects that did not receive (sufficient) attention in earlier work on data selection and domain adaptation:

- The mix-domain parallel corpus C_{mix} contains a large variety of domains that overlap and but also differ in lexical choice and translation. This is radically different from the in-/out-domain setting usually assumed in adaptation and constitutes a major challenge for existing selection approaches.
- The way the small in-domain corpus relates to the large mix-domain corpus is also challenging because translation performance often depends on selecting *relevant* sentence pairs, aside from those that are clearly in-domain.
- The lack of out-domain data in a realistic mix-domain scenario, suggests that efforts are needed at finding data that contrasts enough with the in-domain data. In this work we propose an initial training period (burn-in) for isolating pseudo out-domain data. But it might be that relevance-related approaches could also turn out more effective for this.

In our current model we implement the $P(\mathbf{e} \mid D)$ and $P(\mathbf{f} \mid D)$ as language models, inspired by the approaches based on the contrast between the cross-entropies of in- and mix-domain language models (Moore and Lewis, 2010; Axelrod et al., 2011). However, $P(\mathbf{e} \mid D)$ and $P(\mathbf{f} \mid D)$ should work with *relevance models*, i.e., assessing the relevance of sentences to domain D. *Relevance* is a different concept than fluency as embodied by language models, and this aspects demands special attention in future work.[7]

In ongoing large-scale experiments, we now explore the behavior of our Invitation model on a variety of different data settings and compare that to a range of alternative existing approaches. We are also exploring new variations of our Invitation model to find out what the optimal settings might be for different mixes of domains. So far we find that the burn-in and size of pseudo out-domain selection after burn-in can be important in certain situations. We also observe that estimating the suitable size of the selection set is also a topic that demands more attention because the estimate of $P(D_1)$ with the interpretation *percentage of relevant data in* C_{mix} like likely to demand suitable relevance models instead of language models.

We observe that the present Invitation model could be approached from a discriminative perspective, which could be effective for specific data settings. Finally, it is theoretically not clear whether a single approach will be most effective for all practical data scenarios.

7 Conclusions

This work looks at modeling the relevance of sentence pairs from the mix-domain corpus to a task represented by an in-domain sample. In contrast with previous work we cast this as a translation problem with a latent domain variable. Our *Invitation model* based on iterative weighted Invitations using EM, offers a new view on data selection for MT. Our model also offers principled cut-off points for selecting in-domain and other relevant subsets. Experiments on the in-domain task shows our approach outperforms the existing data selection for such a very complex mixture training data.

[7] We thank Amir Kamran for bringing this difference to our attention through ongoing joint experimental work.

The high accuracy in our experiments in this kind of data compared to the baseline suggests that our model might also offer good estimates that can be used for data weighting. In future work we aim to test the Invitation model for instance weighting and explore avenues for using it for selecting and weighting sub-sentential translation pairs (e.g., phrase pairs) that can be used directly for building SMT systems. A further issue is to improve the quality of word alignments induced for mix-domain corpora. We also aim at exploring a discriminative learning approach in conjunction with our model.

Acknowledgements

The first author is supported by the EXPERT (EXPloiting Empirical appRoaches to Translation) Initial Training Network (ITN) of the European Union's Seventh Framework Programme. We thank Translation Automation Society (TAUS.com) for providing us with suitable data for the mix-domain scenario. We also thank Amir Kamran and Bart Mellebeek for help and collaboration on experiments related to data selection and domain adaptation. We thank Miloš Stanojević and three anonymous reviewers for their valuable comments on earlier versions.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 355–362, Stroudsburg, PA, USA. Association for Computational Linguistics.

Pratyush Banerjee, Sudip Kumar Naskar, Johann Roturier, Andy Way, and Josef van Genabith. 2012. Translation quality-based supplementary data selection by incremental update of translation models. In Martin Kay and Christian Boitet, editors, *COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India*, pages 149–166. Indian Institute of Technology Bombay.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.

Boxing Chen, George Foster, and Roland Kuhn. 2013. Adaptation of reordering models for statistical machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 938–946, Atlanta, Georgia, June. Association for Computational Linguistics.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 427–436, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jonathan H. Clark, 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: Short Papers - Volume 2*, HLT '11, pages 176–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1):1–38.

Michael Denkowski 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*, WMT '11, pages 85–91, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kevin Duh, Graham Neubig, Katsuhito Sudoh, and Hajime Tsukada. 2013. Adaptation data selection using neural language models: Experiments in machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 678–683, Sofia, Bulgaria, August. Association for Computational Linguistics.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 115–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for smt. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 128–135, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Foster, 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*, EMNLP '10, pages 451–459, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. 2012. Does more data always yield better translations? In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 152–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

Barry Haddow and Philipp Koehn. 2012. Analysing the effect of out-of-domain data on smt systems. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, WMT '12, pages 422–432, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ann Irvine, John Morgan, Marine Carpuat, Hal Daume III, and Dragos Munteanu. 2013. Measuring machine translation errors in new domains. *Transactions of the Association for Computational Linguistics (TACL)*.

Philipp Koehn and Barry Haddow. 2012. Towards effective use of training data in statistical machine translation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, Montreal, Canada, June. Association for Computational Linguistics.

Philipp Koehn 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, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Spyros Matsoukas, Antti-Veikko I. Rosti, and Bing Zhang. 2009. Discriminative corpus weight estimation for machine translation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 708–717, Singapore, August. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, pages 220–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Preslav Nakov. 2008. Improving english-spanish statistical machine translation: Experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, StatMT '08, pages 147–150, Stroudsburg, PA, USA. Association for Computational Linguistics.

Radford M. Neal and Geoffrey E. Hinton. 1999. A view of the em algorithm that justifies incremental, sparse, and other variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. MIT Press, Cambridge, MA, USA.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Pauls and Dan Klein. 2011. Faster and smaller n-gram language models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 258–267, Stroudsburg, PA, USA. Association for Computational Linguistics.

William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C (2Nd Ed.): The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA.

Rico Sennrich. 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*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas*, pages 223–231.