

# Active Learning for Multilingual Statistical Machine Translation\*

Gholamreza Haffari and Anoop Sarkar

School of Computing Science, Simon Fraser University

British Columbia, Canada

{ghaffar1,anoop}@cs.sfu.ca

## Abstract

Statistical machine translation (SMT) models require bilingual corpora for training, and these corpora are often multilingual with parallel text in multiple languages simultaneously. We introduce an *active learning* task of adding a new language to an existing multilingual set of parallel text and constructing high quality MT systems, from each language in the collection into this new target language. We show that adding a new language using active learning to the EuroParl corpus provides a significant improvement compared to a random sentence selection baseline. We also provide new highly effective sentence selection methods that improve AL for phrase-based SMT in the multilingual and single language pair setting.

## 1 Introduction

The main source of training data for statistical machine translation (SMT) models is a parallel corpus. In many cases, the same information is available in multiple languages simultaneously as a multilingual parallel corpus, e.g., European Parliament (EuroParl) and U.N. proceedings. In this paper, we consider how to use active learning (AL) in order to add a new language to such a multilingual parallel corpus and at the same time we construct an MT system from each language in the original corpus into this new target language. We introduce a novel combined measure of translation quality for multiple target language outputs (the same content from multiple source languages).

The multilingual setting provides new opportunities for AL over and above a single language pair. This setting is similar to the multi-task AL scenario (Reichart et al., 2008). In our case, the multiple tasks are individual machine translation tasks for several language pairs. The nature of the translation processes vary from any of the source

languages to the new language depending on the characteristics of each source-target language pair, hence these tasks are competing for annotating the same resource. However it may be that in a single language pair, AL would pick a particular sentence for annotation, but in a multilingual setting, a different source language might be able to provide a good translation, thus saving annotation effort. In this paper, we explore how multiple MT systems can be used to effectively pick instances that are more likely to improve training quality.

Active learning is framed as an iterative learning process. In each iteration new human labeled instances (manual translations) are added to the training data based on their expected training quality. However, if we start with only a small amount of initial parallel data for the new target language, then translation quality is very poor and requires a very large injection of human labeled data to be effective. To deal with this, we use a novel framework for active learning: we assume we are given a small amount of parallel text and a large amount of monolingual source language text; using these resources, we create a large noisy parallel text which we then iteratively improve using small injections of human translations. When we build multiple MT systems from multiple source languages to the new target language, each MT system can be seen as a different ‘view’ on the desired output translation. Thus, we can train our multiple MT systems using either *self-training* or *co-training* (Blum and Mitchell, 1998). In self-training each MT system is re-trained using human labeled data plus its own noisy translation output on the unlabeled data. In co-training each MT system is re-trained using human labeled data plus noisy translation output from the other MT systems in the ensemble. We use consensus translations (He et al., 2008; Rosti et al., 2007; Matusov et al., 2006) as an effective method for co-training between multiple MT systems.

This paper makes the following contributions:

- We provide a new framework for multilingual MT, in which we build multiple MT systems and add a new language to an existing multilingual parallel corpus. The multilingual set-

\*Thanks to James Peltier for systems support for our experiments. This research was partially supported by NSERC, Canada (RGPIN: 264905) and an IBM Faculty Award.

ting allows new features for active learning which we exploit to improve translation quality while reducing annotation effort.

- We introduce new highly effective sentence selection methods that improve phrase-based SMT in the multilingual and single language pair setting.
- We describe a novel co-training based active learning framework that exploits consensus translations to effectively select only those sentences that are difficult to translate for all MT systems, thus sharing annotation cost.
- We show that using active learning to add a new language to the EuroParl corpus provides a significant improvement compared to the strong random sentence selection baseline.

## 2 AL-SMT: Multilingual Setting

Consider a multilingual parallel corpus, such as EuroParl, which contains parallel sentences for several languages. Our goal is to add a new language to this corpus, *and* at the same time to construct high quality MT systems from the existing languages (in the multilingual corpus) to the new language. This goal is formalized by the following objective function:

$$\mathcal{O} = \sum_{d=1}^D \alpha_d \times TQ(M_{F^d \rightarrow E}) \quad (1)$$

where  $F^d$ 's are the source languages in the multilingual corpus ( $D$  is the total number of languages), and  $E$  is the new language. The translation quality is measured by  $TQ$  for individual systems  $M_{F^d \rightarrow E}$ ; it can be BLEU score or WER/PER (Word error rate and position independent WER) which induces a maximization or minimization problem, respectively. The non-negative weights  $\alpha_d$  reflect the importance of the different translation tasks and  $\sum_d \alpha_d = 1$ . AL-SMT formulation for *single* language pair is a special case of this formulation where only one of the  $\alpha_d$ 's in the objective function (1) is one and the rest are zero. Moreover the algorithmic framework that we introduce in Sec. 2.1 for AL in the multilingual setting includes the single language pair setting as a special case (Haffari et al., 2009).

We denote the large *unlabeled* multilingual corpus by  $\mathbb{U} := \{(\mathbf{f}_j^1, \dots, \mathbf{f}_j^D)\}$ , and the small *labeled* multilingual corpus by  $\mathbb{L} := \{(\mathbf{f}_i^1, \dots, \mathbf{f}_i^D, \mathbf{e}_i)\}$ . We

overload the term *entry* to denote a tuple in  $\mathbb{L}$  or in  $\mathbb{U}$  (it should be clear from the context). For a single language pair we use  $U$  and  $L$ .

### 2.1 The Algorithmic Framework

Algorithm 1 represents our AL approach for the multilingual setting. We train our initial MT systems  $\{M_{F^d \rightarrow E}\}_{d=1}^D$  on the multilingual corpus  $\mathbb{L}$ , and use them to translate *all* monolingual sentences in  $\mathbb{U}$ . We denote sentences in  $\mathbb{U}$  together with their multiple translations by  $\mathbb{U}^+$  (line 4 of Algorithm 1). Then we retrain the SMT systems on  $\mathbb{L} \cup \mathbb{U}^+$  and use the resulting model to decode the test set. Afterwards, we select and remove a subset of highly informative sentences from  $\mathbb{U}$ , and add those sentences together with their human-provided translations to  $\mathbb{L}$ . This process is continued iteratively until a certain level of translation quality is met (we use the BLEU score, WER and PER) (Papineni et al., 2002). In the baseline, against which we compare our sentence selection methods, the sentences are chosen *randomly*.

When (re-)training the models, two phrase tables are learned for each SMT model: one from the labeled data  $\mathbb{L}$  and the other one from *pseudo-labeled* data  $\mathbb{U}^+$  (which we call the *main* and *auxiliary* phrase tables respectively). (Ueffing et al., 2007; Haffari et al., 2009) show that treating  $\mathbb{U}^+$  as a source for a new feature function in a log-linear model for SMT (Och and Ney, 2004) allows us to maximally take advantage of unlabeled data by finding a weight for this feature using minimum error-rate training (MERT) (Och, 2003).

Since each entry in  $\mathbb{U}^+$  has multiple translations, there are two options when building the auxiliary table for a particular language pair  $(F^d, E)$ : (i) to use the corresponding translation  $\mathbf{e}^d$  of the source language in a *self-training* setting, or (ii) to use the *consensus* translation among all the translation candidates  $(\mathbf{e}^1, \dots, \mathbf{e}^D)$  in a *co-training* setting (sharing information between multiple SMT models).

A whole range of methods exist in the literature for combining the output translations of multiple MT systems for a *single* language pair, operating either at the sentence, phrase, or word level (He et al., 2008; Rosti et al., 2007; Matusov et al., 2006). The method that we use in this work operates at the sentence level, and picks a single high quality translation from the union of the  $n$ -best lists generated by multiple SMT models. Sec. 5 gives

---

**Algorithm 1** AL-SMT-Multiple

---

- 1: Given multilingual corpora  $\mathbb{L}$  and  $\mathbb{U}$
  - 2:  $\{M_{F^d \rightarrow E}\}_{d=1}^D = \mathbf{multtrain}(\mathbb{L}, \emptyset)$
  - 3: **for**  $t = 1, 2, \dots$  **do**
  - 4:  $\mathbb{U}^+ = \mathbf{multtranslate}(\mathbb{U}, \{M_{F^d \rightarrow E}\}_{d=1}^D)$
  - 5: Select  $k$  sentences from  $\mathbb{U}^+$ , and ask a human for their *true* translations.
  - 6: Remove the  $k$  sentences from  $\mathbb{U}$ , and add the  $k$  sentence pairs (translated by human) to  $\mathbb{L}$
  - 7:  $\{M_{F^d \rightarrow E}\}_{d=1}^D = \mathbf{multtrain}(\mathbb{L}, \mathbb{U}^+)$
  - 8: Monitor the performance on the test set
  - 9: **end for**
- 

more details about features which are used in our consensus finding method, and how it is trained. Now let us address the important question of selecting highly informative sentences (step 5 in the Algorithm 1) in the following section.

### 3 Sentence Selection: Multiple Language Pairs

The goal is to optimize the objective function (1) with minimum human effort in providing the translations. This motivates selecting sentences which are *maximally* beneficial for *all* the MT systems. In this section, we present several protocols for sentence selection based on the combined information from multiple language pairs.

#### 3.1 Alternating Selection

The simplest selection protocol is to choose  $k$  sentences (entries) in the first iteration of AL which improve maximally the first model  $M_{F^1 \rightarrow E}$ , while ignoring other models. In the second iteration, the sentences are selected with respect to the second model, and so on (Reichart et al., 2008).

#### 3.2 Combined Ranking

Pick any AL-SMT scoring method for a *single* language pair (see Sec. 4). Using this method, we rank the entries in unlabeled data  $\mathbb{U}$  for each translation task defined by language pair  $(F^d, E)$ . This results in several ranking lists, each of which represents the importance of entries with respect to a particular translation task. We combine these rankings using a combined score:

$$\text{Score}((\mathbf{f}^1, \dots, \mathbf{f}^D)) = \sum_{d=1}^D \alpha_d \text{Rank}_d(\mathbf{f}^d)$$

$\text{Rank}_d(\cdot)$  is the ranking of a sentence in the list for the  $d^{\text{th}}$  translation task (Reichart et al., 2008).

### 3.3 Disagreement Among the Translations

Disagreement among the candidate translations of a particular entry is evidence for the difficulty of that entry for different translation models. The reason is that disagreement increases the possibility that most of the translations are not correct. Therefore it would be beneficial to ask human for the translation of these hard entries.

Now the question is how to quantify the notion of disagreement among the candidate translations  $(\mathbf{e}^1, \dots, \mathbf{e}^D)$ . We propose two measures of disagreement which are related to the portion of shared  $n$ -grams ( $n \leq 4$ ) among the translations:

- Let  $\mathbf{e}^c$  be the consensus among all the candidate translations, then define the disagreement as  $\sum_d \alpha_d (1 - \text{BLEU}(\mathbf{e}^c, \mathbf{e}^d))$ .
- Based on the disagreement of every pair of candidate translations:  $\sum_d \alpha_d \sum_{d'} (1 - \text{BLEU}(\mathbf{e}^{d'}, \mathbf{e}^d))$ .

For the single language pair setting, (Haffari et al., 2009) presents and compares several sentence selection methods for statistical phrase-based machine translation. We introduce novel techniques which outperform those methods in the next section.

### 4 Sentence Selection: Single Language Pair

Phrases are basic units of translation in phrase-based SMT models. The phrases which may potentially be extracted from a sentence indicate its informativeness. The more new phrases a sentence can offer, the more informative it is; since it boosts the *generalization* of the model. Additionally phrase translation probabilities need to be estimated accurately, which means sentences that offer phrases whose occurrences in the corpus were rare are informative. When selecting new sentences for human translation, we need to pay attention to this tradeoff between *exploration* and *exploitation*, i.e. selecting sentences to discover new phrases v.s. estimating accurately the phrase translation probabilities. Smoothing techniques partly handle accurate estimation of translation probabilities when the events occur rarely (indeed it is the main reason for smoothing). So we mainly focus on how to expand effectively the lexicon or set of phrases of the model.

The more frequent a phrase (not a *phrase pair*) is in the *unlabeled* data, the more important it is to

know its translation; since it is more likely to see it in test data (specially when the test data is in-domain with respect to unlabeled data). The more frequent a phrase is in the *labeled* data, the more unimportant it is; since probably we have observed most of its translations.

In the labeled data  $L$ , phrases are the ones which are extracted by the SMT models; but what are the candidate phrases in the unlabeled data  $U$ ? We use the currently trained SMT models to answer this question. Each translation in the  $n$ -best list of translations (generated by the SMT models) corresponds to a particular segmentation of a sentence, which breaks that sentence into several fragments (see Fig. 1). Some of these fragments are the source language part of a phrase pair available in the phrase table, which we call *regular phrases* and denote their set by  $X_s^{reg}$  for a sentence  $s$ . However, there are some fragments in the sentence which are *not* covered by the phrase table – possibly because of the OOVs (out-of-vocabulary words) or the constraints imposed by the phrase extraction algorithm – called  $X_s^{oov}$  for a sentence  $s$ . Each member of  $X_s^{oov}$  offers a set of *potential phrases* (also referred to as *OOV phrases*) which are not observed due to the *latent* segmentation of this fragment. We present two generative models for the phrases and show how to estimate and use them for sentence selection.

#### 4.1 Model 1

In the first model, the generative story is to generate phrases for each sentence based on independent draws from a multinomial. The sample space of the multinomial consists of both regular and OOV phrases.

We build two models, i.e. two multinomials, one for labeled data and the other one for unlabeled data. Each model is trained by maximizing the log-likelihood of its corresponding data:

$$\mathcal{L}_{\mathcal{D}} := \sum_{s \in \mathcal{D}} \tilde{P}(s) \sum_{\mathbf{x} \in X_s} \log P(\mathbf{x} | \boldsymbol{\theta}_{\mathcal{D}}) \quad (2)$$

where  $\mathcal{D}$  is either  $L$  or  $U$ ,  $\tilde{P}(s)$  is the *empirical* distribution of the sentences<sup>1</sup>, and  $\boldsymbol{\theta}_{\mathcal{D}}$  is the parameter vector of the corresponding probability

<sup>1</sup> $\tilde{P}(s)$  is the number of times that the sentence  $s$  is seen in  $\mathcal{D}$  divided by the number of all sentences in  $\mathcal{D}$ .

distribution. When  $\mathbf{x} \in X_s^{oov}$ , we will have

$$\begin{aligned} P(\mathbf{x} | \boldsymbol{\theta}_U) &= \sum_{h \in H_{\mathbf{x}}} P(\mathbf{x}, h | \boldsymbol{\theta}_U) \\ &= \sum_{h \in H_{\mathbf{x}}} P(h) P(\mathbf{x} | h, \boldsymbol{\theta}_U) \\ &= \frac{1}{|H_{\mathbf{x}}|} \sum_{h \in H_{\mathbf{x}}} \prod_{\mathbf{y} \in Y_{\mathbf{x}}^h} \boldsymbol{\theta}_U(\mathbf{y}) \quad (3) \end{aligned}$$

where  $H_{\mathbf{x}}$  is the space of all possible segmentations for the OOV fragment  $\mathbf{x}$ ,  $Y_{\mathbf{x}}^h$  is the resulting phrases from  $\mathbf{x}$  based on the segmentation  $h$ , and  $\boldsymbol{\theta}_U(\mathbf{y})$  is the probability of the OOV phrase  $\mathbf{y}$  in the multinomial associated with  $U$ . We let  $H_{\mathbf{x}}$  to be all possible segmentations of the fragment  $\mathbf{x}$  for which the resulting phrase lengths are not greater than the maximum length constraint for phrase extraction in the underlying SMT model. Since we do not know anything about the segmentations a priori, we have put a uniform distribution over such segmentations.

Maximizing (2) to find the maximum likelihood parameters for this model is an extremely difficult problem<sup>2</sup>. Therefore, we maximize the following lower-bound on the log-likelihood which is derived using Jensen’s inequality:

$$\begin{aligned} \mathcal{L}_{\mathcal{D}} &\geq \sum_{s \in \mathcal{D}} \tilde{P}(s) \left[ \sum_{\mathbf{x} \in X_s^{reg}} \log \boldsymbol{\theta}_{\mathcal{D}}(\mathbf{x}) \right. \\ &\quad \left. + \sum_{\mathbf{x} \in X_s^{oov}} \sum_{h \in H_{\mathbf{x}}} \frac{1}{|H_{\mathbf{x}}|} \sum_{\mathbf{y} \in Y_{\mathbf{x}}^h} \log \boldsymbol{\theta}_{\mathcal{D}}(\mathbf{y}) \right] \quad (4) \end{aligned}$$

Maximizing (4) amounts to set the probability of each regular / potential phrase proportional to its count / *expected* count in the data  $\mathcal{D}$ .

Let  $\rho_k(\mathbf{x}_{i:j})$  be the number of possible segmentations from position  $i$  to position  $j$  of an OOV fragment  $\mathbf{x}$ , and  $k$  is the maximum phrase length;

$$\rho_k(\mathbf{x}_{1:|\mathbf{x}|}) = \begin{cases} 0, & \text{if } |\mathbf{x}| = 0 \\ 1, & \text{if } |\mathbf{x}| = 1 \\ \sum_{i=1}^k \rho_k(\mathbf{x}_{i+1:|\mathbf{x}|}), & \text{otherwise} \end{cases}$$

which gives us a dynamic programming algorithm to compute the number of segmentation  $|H_{\mathbf{x}}| = \rho_k(\mathbf{x}_{1:|\mathbf{x}|})$  of the OOV fragment  $\mathbf{x}$ . The expected count of a potential phrase  $\mathbf{y}$  based on an OOV segment  $\mathbf{x}$  is (see Fig. 1.c):

$$E[\mathbf{y} | \mathbf{x}] = \frac{\sum_{i \leq j} \delta_{[\mathbf{y} = \mathbf{x}_{i:j}]} \rho_k(\mathbf{x}_{1:i-1}) \rho_k(\mathbf{x}_{j+1:|\mathbf{x}|})}{\rho_k(\mathbf{x})}$$

<sup>2</sup>Setting partial derivatives of the Lagrangian to zero amounts to finding the roots of a system of multivariate polynomials (a major topic in Algebraic Geometry).

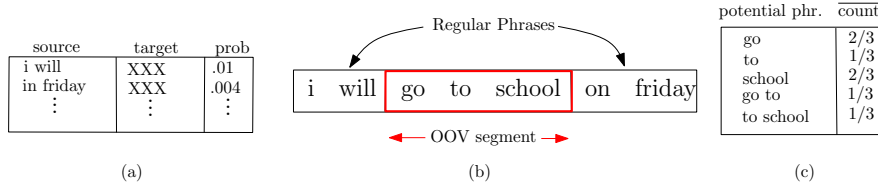


Figure 1: The given sentence in (b) is segmented, based on the source side phrases extracted from the phrase table in (a), to yield regular phrases and OOV segment. The table in (c) shows the potential phrases extracted from the OOV segment “go to school” and their expected counts (denoted by  $\overline{\text{count}}$ ) where the maximum length for the potential phrases is set to 2. In the example, “go to school” has 3 segmentations with maximum phrase length 2:  $(go)(to\ school)$ ,  $(go\ to)(school)$ ,  $(go)(to)(school)$ .

where  $\delta_{[C]}$  is 1 if the condition  $C$  is true, and zero otherwise. We have used the fact that the number of occurrences of a phrase spanning the indices  $[i, j]$  is the product of the number of segmentations of the left and the right sub-fragments, which are  $\rho_k(\mathbf{x}_{1:i-1})$  and  $\rho_k(\mathbf{x}_{j+1:|\mathbf{x}|})$  respectively.

## 4.2 Model 2

In the second model, we consider a mixture model of two multinomials responsible for generating phrases in each of the labeled and unlabeled data sets. To generate a phrase, we first toss a coin and depending on the outcome we either generate the phrase from the multinomial associated with regular phrases  $\theta_U^{reg}$  or potential phrases  $\theta_U^{oov}$ :

$$P(\mathbf{x}|\theta_U) := \beta_U \theta_U^{reg}(\mathbf{x}) + (1 - \beta_U) \theta_U^{oov}(\mathbf{x})$$

where  $\theta_U$  includes the mixing weight  $\beta$  and the parameter vectors of the two multinomials. The mixture model associated with  $L$  is written similarly. The parameter estimation is based on maximizing a lower-bound on the log-likelihood which is similar to what was done for the Model 1.

## 4.3 Sentence Scoring

The sentence score is a linear combination of two terms: one coming from regular phrases and the other from OOV phrases:

$$\begin{aligned} \phi_1(\mathbf{s}) := & \frac{\lambda}{|X_s^{reg}|} \sum_{\mathbf{x} \in X_s^{reg}} \log \frac{P(\mathbf{x}|\theta_U)}{P(\mathbf{x}|\theta_L)} \\ & + \frac{1 - \lambda}{|X_s^{oov}|} \sum_{\mathbf{x} \in X_s^{oov}} \sum_{h \in H_{\mathbf{x}}} \frac{1}{|H_{\mathbf{x}}|} \log \prod_{\mathbf{y} \in Y_{\mathbf{x}}^h} \frac{P(\mathbf{y}|\theta_U)}{P(\mathbf{y}|\theta_L)} \end{aligned}$$

where we use either Model 1 or Model 2 for  $P(\cdot|\theta_D)$ . The first term is the log probability ratio of regular phrases under phrase models corresponding to unlabeled and labeled data, and the second term is the *expected log probability ratio* (ELPR) under the two models. Another option for

the contribution of OOV phrases is to take log of *expected probability ratio* (LEPR):

$$\begin{aligned} \phi_2(\mathbf{s}) := & \frac{\lambda}{|X_s^{reg}|} \sum_{\mathbf{x} \in X_s^{reg}} \log \frac{P(\mathbf{x}|\theta_U)}{P(\mathbf{x}|\theta_L)} \\ & + \frac{1 - \lambda}{|X_s^{oov}|} \sum_{\mathbf{x} \in X_s^{oov}} \log \sum_{h \in H_{\mathbf{x}}} \frac{1}{|H_{\mathbf{x}}|} \prod_{\mathbf{y} \in Y_{\mathbf{x}}^h} \frac{P(\mathbf{y}|\theta_U)}{P(\mathbf{y}|\theta_L)} \end{aligned}$$

It is not difficult to prove that there is no difference between Model 1 and Model 2 when ELPR scoring is used for sentence selection. However, the situation is different for LEPR scoring: the two models produce different sentence rankings in this case.

## 5 Experiments

**Corpora.** We pre-processed the EuroParl corpus (<http://www.statmt.org/europarl>) (Koehn, 2005) and built a multilingual parallel corpus with 653,513 sentences, excluding the Q4/2000 portion of the data (2000-10 to 2000-12) which is reserved as the test set. We subsampled 5,000 sentences as the labeled data  $\mathbb{L}$  and 20,000 sentences as  $\mathbb{U}$  for the pool of untranslated sentences (while hiding the English part). The test set consists of 2,000 multi-language sentences and comes from the multilingual parallel corpus built from Q4/2000 portion of the data.

**Consensus Finding.** Let  $T$  be the union of the  $n$ -best lists of translations for a particular sentence. The consensus translation  $\mathbf{t}^c$  is

$$\arg \max_{\mathbf{t} \in T} w_1 \frac{LM(\mathbf{t})}{|\mathbf{t}|} + w_2 \frac{Q_d(\mathbf{t})}{|\mathbf{t}|} + w_3 R_d(\mathbf{t}) + w_{4,d}$$

where  $LM(\mathbf{t})$  is the score from a 3-gram language model,  $Q_d(\mathbf{t})$  is the translation score generated by the decoder for  $M_{F^d \rightarrow E}$  if  $\mathbf{t}$  is produced by the  $d$ th SMT model,  $R_d(\mathbf{t})$  is the rank of the translation in the  $n$ -best list produced by the  $d$ th model,  $w_{4,d}$  is a bias term for each translation model to make their scores comparable, and  $|\mathbf{t}|$  is the length

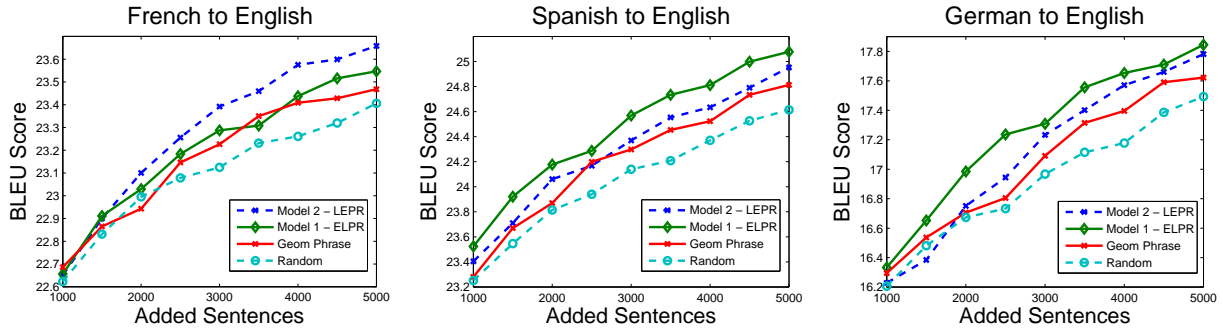


Figure 2: The performance of different sentence selection methods as the iteration of AL loop goes on for three translation tasks. Plots show the performance of sentence selection methods for single language pair in Sec. 4 compared to the GeomPhrase (Haffari et al., 2009) and random sentence selection baseline.

of the translation sentence. The number of weights  $w_i$  is 3 plus the number of source languages, and they are trained using minimum error-rate training (MERT) to maximize the BLEU score (Och, 2003) on a development set.

**Parameters.** We use add- $\epsilon$  smoothing where  $\epsilon = .5$  to smooth the probabilities in Sec. 4; moreover  $\lambda = .4$  for ELPR and LEPR sentence scoring and maximum phrase length  $k$  is set to 4. For the multilingual experiments (which involve four source languages) we set  $\alpha_d = .25$  to make the importance of individual translation tasks equal.

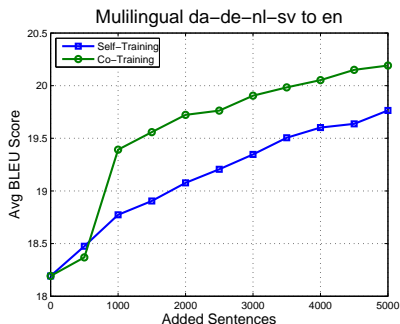


Figure 3: Random sentence selection baseline using self-training and co-training (Germanic languages to English).

## 5.1 Results

First we evaluate the proposed sentence selection methods in Sec. 4 for the single language pair. Then the best method from the single language pair setting is used to evaluate sentence selection methods for AL in multilingual setting. After building the initial MT system for each experiment, we select and remove 500 sentences from  $\mathbb{U}$  and add them together with translations to  $\mathbb{L}$  for 10 total iterations. The random sentence selection baselines are averaged over 3 independent runs.

mode Method	self-train		co-train	
	wer	per	wer	per
Combined Rank	<b>40.2</b>	<b>30.0</b>	<b>40.0</b>	<b>29.6</b>
Alternate	41.0	30.2	40.1	30.1
Disagree-Pairwise	41.9	32.0	40.5	30.9
Disagree-Center	41.8	31.8	40.6	30.7
Random Baseline	41.6	31.0	40.5	30.7

Germanic languages to English

mode Method	self-train		co-train	
	wer	per	wer	per
Combined Rank	<b>37.7</b>	<b>27.3</b>	<b>37.3</b>	<b>27.0</b>
Alternate	<b>37.7</b>	<b>27.3</b>	<b>37.3</b>	<b>27.0</b>
Random Baseline	38.6	28.1	38.1	27.6

Romance languages to English

Table 1: Comparison of multilingual selection methods with WER (word error rate), PER (position independent WER). 95% confidence interval for WER numbers is 0.7 and for PER numbers is 0.5. **Bold**: best result, *italic*: significantly better.

We use three language pairs in our single language pair experiments: French-English, German-English, and Spanish-English. In addition to random sentence selection baseline, we also compare the methods proposed in this paper to the best method reported in (Haffari et al., 2009) denoted by GeomPhrase, which differs from our models since it considers each individual OOV segment as a single OOV phrase and does not consider subsequences. The results are presented in Fig. 2. Selecting sentences based on our proposed methods outperform the random sentence selection baseline and GeomPhrase. We suspect for the situations where  $\mathbb{L}$  is out-of-domain and the average phrase length is relatively small, our method will outperform GeomPhrase even more.

For the multilingual experiments, we use Germanic (German, Dutch, Danish, Swedish) and Romance (French, Spanish, Italian, Portuguese<sup>3</sup>) lan-

<sup>3</sup>A reviewer pointed out that EuroParl English-Portuguese

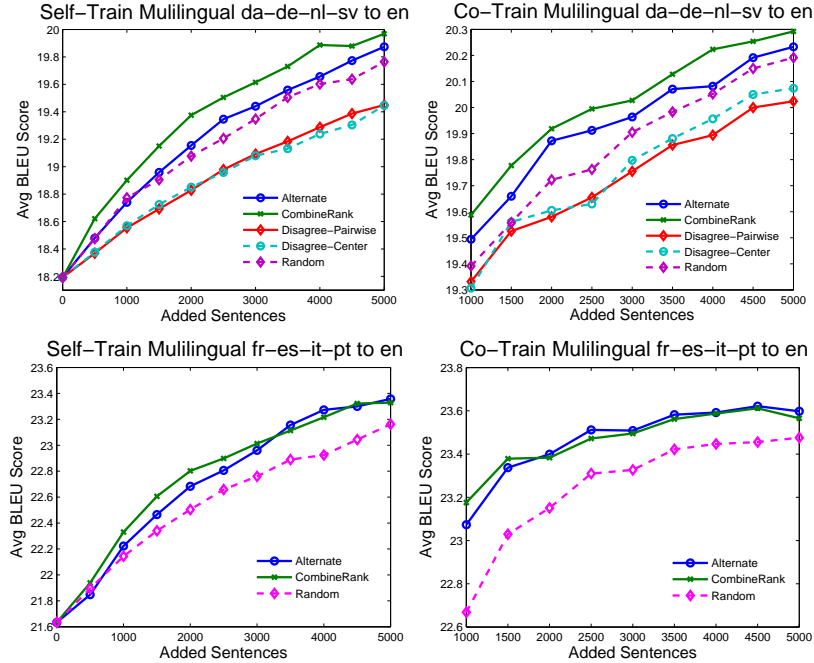


Figure 4: The left/right plot show the performance of our AL methods for multilingual setting combined with self-training/co-training. The sentence selection methods from Sec. 3 are compared with random sentence selection baseline. The top plots correspond to Danish-German-Dutch-Swedish to English, and the bottom plots correspond to French-Spanish-Italian-Portuguese to English.

languages as the source and English as the target language as two sets of experiments.<sup>4</sup> Fig. 3 shows the performance of random sentence selection for AL combined with self-training/co-training for the multi-source translation from the four Germanic languages to English. It shows that the co-training mode outperforms the self-training mode by almost 1 BLEU point. The results of selection strategies in the multilingual setting are presented in Fig. 4 and Tbl. 1. Having noticed that Model 1 with ELPR performs well in the single language pair setting, we use it to rank entries for individual translation tasks. Then these rankings are used by ‘Alternate’ and ‘Combined Rank’ selection strategies in the multilingual case. The ‘Combined Rank’ method outperforms all the other methods including the strong random selection baseline in both self-training and co-training modes. The disagreement-based selection methods underperform the baseline for translation of Germanic languages to English, so we omitted them for the Romance language experiments.

## 5.2 Analysis

The basis for our proposed methods has been the popularity of regular/OOV phrases in  $U$  and their data is very noisy and future work should omit this pair.

<sup>4</sup>Choice of Germanic and Romance for our experimental setting is inspired by results in (Cohn and Lapata, 2007)

unpopularity in  $L$ , which is measured by  $\frac{P(\mathbf{x}|\theta_U)}{P(\mathbf{x}|\theta_L)}$ . We need  $P(\mathbf{x}|\theta_U)$ , the *estimated* distribution of phrases in  $U$ , to be as similar as possible to  $P^*(\mathbf{x})$ , the *true* distribution of phrases in  $U$ . We investigate this issue for regular/OOV phrases as follows:

- Using the output of the initially trained MT system on  $L$ , we extract the regular/OOV phrases as described in §4. The smoothed relative frequencies give us the regular/OOV phrasal distributions.
- Using the *true* English translation of the sentences in  $U$ , we extract the true phrases. Separating the phrases into two sets of regular and OOV phrases defined by the previous step, we use the smoothed relative frequencies and form the *true* OOV/regular phrasal distributions.

We use the KL-divergence to see how dissimilar are a pair of given probability distributions. As Tbl. 2 shows, the KL-divergence between the true and estimated distributions are less than that

	De2En	Fr2En	Es2En
$KL(P_{reg}^*    P_{reg})$	4.37	4.17	4.38
$KL(P_{reg}^*    unif)$	5.37	5.21	5.80
$KL(P_{oov}^*    P_{oov})$	3.04	4.58	4.73
$KL(P_{oov}^*    unif)$	3.41	4.75	4.99

Table 2: For regular/OOV phrases, the KL-divergence between the true distribution ( $P^*$ ) and the estimated ( $P$ ) or uniform ( $unif$ ) distributions are shown, where:

$$KL(P^* || P) := \sum_x P^*(x) \log \frac{P^*(x)}{P(x)}.$$

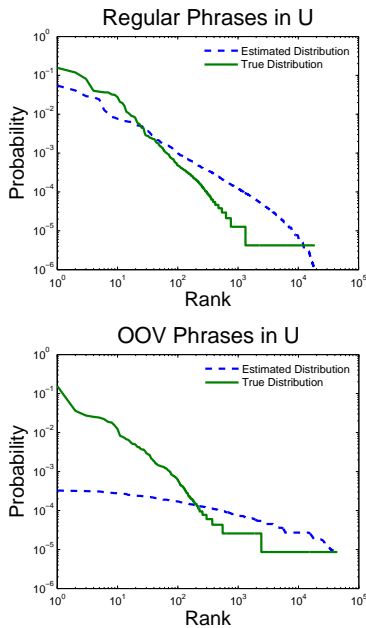


Figure 5: The log-log Zipf plots representing the true and estimated probabilities of a (source) *phrase* vs the rank of that phrase in the German to English translation task. The plots for the Spanish to English and French to English tasks are also similar to the above plots, and confirm a power law behavior in the true phrasal distributions.

between the true and uniform distributions, in all three language pairs. Since uniform distribution conveys no information, this is evidence that there is some information encoded in the estimated distribution about the true distribution. However we noticed that the true distributions of regular/OOV phrases exhibit Zipfian (power law) behavior<sup>5</sup> which is not well captured by the estimated distributions (see Fig. 5). Enhancing the estimated distributions to capture this power law behavior would improve the quality of the proposed sentence selection methods.

## 6 Related Work

(Haffari et al., 2009) provides results for active learning for MT using a single language pair. Our work generalizes to the use of multilingual corpora using new methods that are not possible with a single language pair. In this paper, we also introduce new selection methods that outperform the methods in (Haffari et al., 2009) even for MT with a single language pair. In addition in this paper by considering multilingual parallel corpora we were able to introduce co-training for AL, while (Haffari et al., 2009) only use self-training since they are using a single language pair.

<sup>5</sup>This observation is at the *phrase* level and not at the *word* (Zipf, 1932) or even *n*-gram level (Ha et al., 2002).

(Reichart et al., 2008) introduces multi-task active learning where unlabeled data require annotations for multiple tasks, e.g. they consider named entities and parse trees, and showed that multiple tasks helps selection compared to individual tasks. Our setting is different in that the target language is the same across multiple MT tasks, which we exploit to use consensus translations and co-training to improve active learning performance.

(Callison-Burch and Osborne, 2003b; Callison-Burch and Osborne, 2003a) provide a co-training approach to MT, where one language pair creates data for another language pair. In contrast, our co-training approach uses consensus translations and our setting for active learning is very different from their semi-supervised setting. A Ph.D. proposal by Chris Callison-Burch (Callison-burch, 2003) lays out the promise of AL for SMT and proposes some algorithms. However, the lack of experimental results means that performance and feasibility of those methods cannot be compared to ours.

While we use consensus translations (He et al., 2008; Rosti et al., 2007; Matusov et al., 2006) as an effective method for co-training in this paper, unlike consensus for system combination, the source languages for each of our MT systems are different, which rules out a set of popular methods for obtaining consensus translations which assume translation for a single language pair. Finally, we briefly note that triangulation (see (Cohn and Lapata, 2007)) is orthogonal to the use of co-training in our work, since it only enhances each MT system in our ensemble by exploiting the multilingual data. In future work, we plan to incorporate triangulation into our active learning approach.

## 7 Conclusion

This paper introduced the novel active learning task of adding a new language to an existing multilingual set of parallel text. We construct SMT systems from each language in the collection into the new target language. We show that we can take advantage of multilingual corpora to decrease annotation effort thanks to the highly effective sentence selection methods we devised for active learning in the single language-pair setting which we then applied to the multilingual sentence selection protocols. In the multilingual setting, a novel co-training method for active learning in SMT is proposed using consensus translations which outperforms AL-SMT with self-training.

## References

- Avrim Blum and Tom Mitchell. 1998. Combining Labeled and Unlabeled Data with Co-Training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT 1998)*, Madison, Wisconsin, USA, July 24-26. ACM.
- Chris Callison-Burch and Miles Osborne. 2003a. Bootstrapping parallel corpora. In *NAACL workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*.
- Chris Callison-Burch and Miles Osborne. 2003b. Co-training for statistical machine translation. In *Proceedings of the 6th Annual CLUK Research Colloquium*.
- Chris Callison-burch. 2003. Active learning for statistical machine translation. In *PhD Proposal, Edinburgh University*.
- Trevor Cohn and Mirella Lapata. 2007. Machine translation by triangulation: Making effective use of multi-parallel corpora. In *ACL*.
- Le Quan Ha, E. I. Sicilia-Garcia, Ji Ming, and F.J. Smith. 2002. Extension of zipf's law to words and phrases. In *Proceedings of the 19th international conference on Computational linguistics*.
- Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *NAACL*.
- Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen, and Robert Moore. 2008. Indirect-hmm-based hypothesis alignment for combining outputs from machine translation systems. In *EMNLP*.
- Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.
- Evgeny Matusov, Nicola Ueffing, and Hermann Ney. 2006. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *EACL*.
- Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417-449.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*.
- Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *ACL*.
- Antti-Veikko Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard M. Schwartz, and Bonnie Jean Dorr. 2007. Combining outputs from multiple machine translation systems. In *NAACL*.
- Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *ACL*.
- George Zipf. 1932. *Selective Studies and the Principle of Relative Frequency in Language*. Harvard University Press.