Active Learning for Statistical Phrase-based Machine Translation*

Gholamreza Haffari and **Maxim Roy** and **Anoop Sarkar**
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
{ghaffar1,maximr,anoop}@cs.sfu.ca

Abstract

Statistical machine translation (SMT) models need large bilingual corpora for training, which are unavailable for some language pairs. This paper provides the first serious experimental study of active learning for SMT. We use active learning to improve the quality of a phrase-based SMT system, and show significant improvements in translation compared to a random sentence selection baseline, when test and training data are taken from the same or different domains. Experimental results are shown in a simulated setting using three language pairs, and in a realistic situation for Bangla-English, a language pair with limited translation resources.

1 Introduction

Statistical machine translation (SMT) systems have made great strides in translation quality. However, high quality translation output is dependent on the availability of massive amounts of parallel text in the source and target language. However, there are a large number of languages that are considered "low-density", either because the population speaking the language is not very large, or even if millions of people speak the language, insufficient amounts of parallel text are available in that language.

A statistical translation system can be improved or adapted by incorporating new training data in the form of parallel text. In this paper, we propose several novel *active learning* (AL) strategies for statistical machine translation in order to attack this problem. Conventional techniques for AL of classifiers are problematic in the SMT setting. Selective sampling of sentences for AL may lead to a parallel corpus where each sentence does not share any phrase pairs with the others. Thus, new sentences cannot be translated since we lack evidence for how phrase pairs combine to form novel translations. In this paper, we take the approach of exploration vs. exploitation: where in some cases we pick sentences that are not entirely novel to improve translation statistics, while also injecting novel translation pairs to improve coverage.

There may be evidence to show that AL is useful even when we have massive amounts of parallel training data. (Turchi et al., 2008) presents a comprehensive learning curve analysis of a phrase-based SMT system, and one of the conclusions they draw is, "The first obvious approach is an effort to identify or produce data sets on demand (active learning, where the learning system can request translations of specific sentences, to satisfy its information needs)."

Despite the promise of active learning for SMT there has been very little experimental work published on this issue (see Sec. 5). In this paper, we make several novel contributions to the area of active learning for SMT:

• We use a novel framework for AL, which to our knowledge has not been used in AL experiments before. We assume 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.

• We provide many useful and novel features useful for AL in SMT. In translation, we can leverage a whole new set of features that were out of reach for classification systems: we devise features that look at the source language, but also devise features that make an estimate of the potential utility of translations from the source, e.g. phrase pairs that could be extracted.

• We show that AL can be useful in domain adaptation. We provide the first experimental evidence in SMT that active learning can be used to inject care-

*We would like to thank Chris Callison-Burch for fruitful discussions. This research was partially supported by NSERC, Canada (RGPIN: 264905) and by an IBM Faculty Award to the third author.

fully selected translations in order to improve SMT output in a new domain.

• We compare our proposed features to a random selection baseline in a simulated setting for three language pairs. We also use a realistic setting: using human expert annotations in our AL system we create an improved SMT system to translate from Bangla to English, a language pair with very few resources.

2 An Active Learning Framework for SMT

Starting from an SMT model trained initially on bilingual data, the problem is to minimize the human effort in translating new sentences which will be added to the training data to make the *retrained* SMT model achieves a certain level of performance. Thus, given a bitext $L := \{(\mathbf{f}_i, \mathbf{e}_i)\}$ and a monolingual source text $U := \{\mathbf{f}_j\}$, the goal is to select a subset of highly informative sentences from U to present to a human expert for translation. Highly informative sentences are those which, together with their translations, help the retrained SMT system *quickly* reach a certain level of translation quality. This learning scenario is known as active learning with Selective Sampling (Cohn et al., 1994).

Algorithm 1 describes the experimental setup we propose for active learning. We train our initial MT system on the bilingual corpus L, and use it to translate *all* monolingual sentences in U. We denote sentences in U together with their translations as U^+ (line 4 of Algorithm 1). Then we retrain the SMT system on $L \cup U^+$ and use the resulting model to decode the test set. Afterwards, we select and remove a subset of highly informative sentences from U, and add those sentences together with their human-provided translations to L. This process is continued iteratively until a certain level of translation quality, which in our case is measured by the BLEU score, is met. In the baseline, against which we compare our sentence selection methods, the sentences are chosen *randomly*.

When (re-)training the model, two phrase tables are learned: one from L and the other one from U^+. The phrase table obtained from U^+ is added as a new feature function in the log-linear translation model. The alternative is to ignore U^+ as in a conventional AL setting, however, in our experiments we have found that using more bilingual data, even noisy data, results in better translations.

Algorithm 1 AL-SMT

1: Given bilingual corpus L, and monolingual corpus U.
2: $M_{F \to E} = \textbf{train}(L, \emptyset)$
3: **for** $t = 1, 2, ...$ **do**
4: $\quad U^+ = \textbf{translate}(U, M_{F \to E})$
5: $\quad$ Select k sentence pairs from U^+, and ask a human for their *true* translations.
6: $\quad$ Remove the k sentences from U, and add the k sentence pairs (translated by human) to L
7: $\quad M_{F \to E} = \textbf{train}(L, U^+)$
8: $\quad$ Monitor the performance on the test set T
9: **end for**

Phrase tables from U^+ will get a 0 score in minimum error rate training if they are not useful, so our method is more general. Also, this method has been shown empirically to be more effective (Ueffing et al., 2007b) than (1) using the weighted combination of the two phrase tables from L and U^+, or (2) combining the two sets of data and training from the bitext $L \cup U^+$.

The setup in Algorithm 1 helps us to investigate how to maximally take advantage of human effort (for sentence translation) when learning an SMT model from the available data, that includes bilingual and monolingual text.

3 Sentence Selection Strategies

Our sentence selection strategies can be divided into two categories: (1) those which are independent of the target language and just look into the source language, and (2) those which also take into account the target language. From the description of the methods, it will be clear to which category they belong to. We will see in Sec. 4 that the most promising sentence selection strategies belong to the second category.

3.1 The Utility of Translation Units

Phrases are basic units of translation in phrase-based SMT models. The phrases potentially extracted from a sentence indicate its informativeness. The more new phrases a sentence can offer, the more informative it is. Additionally phrase translation probabilities need to be estimated accurately, which means sentences that contain rare phrases are also informative. When selecting new sentences for hu-

man translation, we need to pay attention to this tradeoff between *exploration* and *exploitation*, i.e. selecting sentences to discover new phrases vs estimating accurately the phrase translation probabilities. A similar argument can be made that emphasizes the importance of words rather than phrases for any SMT model. Also we should take into account that smoothing is a means for accurate estimation of translation probabilities when events are rare. In our work, we focus on methods that effectively expand the lexicon or set of phrases of the model.

3.1.1 Phrases (Geom-Phrase, Arith-Phrase)[1]

The more frequent a phrase is in the *unlabeled* data, the more important it is to know its translation; since it is more likely to occur in the test data (especially 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.

Based on the above observations, we measure the importance score of a sentence as:

$$\phi_g^p(s) := \Big[\prod_{x \in X_s^p} \frac{P(x|U)}{P(x|L)} \Big]^{\frac{1}{|X_s^p|}} \quad (1)$$

where X_s^p is the set of possible phrases that sentence s can offer, and $P(x|\mathcal{D})$ is the probability of observing x in the data $\mathcal{D}$: $P(x|\mathcal{D}) = \frac{Count(x)+\epsilon}{\sum_{x \in X_{\mathcal{D}}^p} Count(x)+\epsilon}$. The score (1) is the averaged *probability ratio* of the set of candidate phrases, i.e. the probability of the candidate phrases under a probabilistic phrase model based on U divided by that based on L. In addition to the geometric average in (1), we may also consider the arithmetic average score:

$$\phi_a^p(s) := \frac{1}{|X_s^p|} \sum_{x \in X_s^p} \frac{P(x|U)}{P(x|L)} \quad (2)$$

Note that (1) can be re-written as $\frac{1}{|X_s^p|} \sum_{x \in X_s^p} \log \frac{P(x|U)}{P(x|L)}$ in the logarithm space, which is similar to (2) with the difference of additional $\log$.

In parallel data L, phrases are the ones which are extracted by the usual phrase extraction algorithm; but what are the candidate phrases in the unlabeled data? Considering the k-best list of translations can tell us the possible phrases the input sentence may offer. For each translation, we have access to the phrases used by the decoder to produce that output. However, there may be islands of out-of-vocabulary (OOV) words that were not in the phrase table and not translated by the decoder as a phrase. We group together such groups of OOV words to form an OOV phrase. The set of possible phrases we extract from the decoder output contain those coming from the phrase table (from labeled data L) and those coming from OOVs. OOV phrases are also used in our computation, where $P(x \mid L)$ for an OOV phrase x is the uniform probability over all OOV phrases.

3.1.2 n-grams (Geom n-gram, Arith n-gram)

As an alternative to phrases, we consider n-grams as basic units of generalization. The resulting score is the weighted combination of the n-gram based scores:

$$\phi_g^N(s) := \sum_{n=1}^{N} \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \log \frac{P(x|U,n)}{P(x|L,n)} \quad (3)$$

where X_s^n denotes n-grams in the sentence s, and $P(x|\mathcal{D},n)$ is the probability of x in the set of n-grams in $\mathcal{D}$. The weights w_n adjust the importance of the scores of n-grams with different lengths. In addition to taking geometric average, we also consider the arithmetic average:

$$\phi_a^N(s) := \sum_{n=1}^{N} \frac{w_n}{|X_s^n|} \sum_{x \in X_s^n} \frac{P(x|U,n)}{P(x|L,n)} \quad (4)$$

As a special case when $N = 1$, the score motivates selecting sentences which increase the number of unique words with new words appearing with higher frequency in U than L.

3.2 Similarity to the Bilingual Training Data (Similarity)

The simplest way to expand the lexicon set is to choose sentences from U which are as dissimilar as possible to L. We measure the similarity using weighted n-gram coverage (Ueffing et al., 2007b).

3.3 Confidence of Translations (Confidence)

The decoder produces an output translation $\mathbf{e}$ using the probability $p(\mathbf{e} \mid \mathbf{f})$. This probability can be

[1] The names in the parentheses are short names used to identify the method in the experimental results.

treated as a confidence score for the translation. To make the confidence score for sentences with different lengths comparable, we normalize using the sentence length (Ueffing et al., 2007b).

3.4 Feature Combination (Combined)

The idea is to take into account the information from several simpler methods, e.g. those mentioned in Sec. 3.1–3.3, when producing the final ranking of sentences. We can either merge the output rankings of those simpler models[2], or use the scores generated by them as input *features* for a higher level ranking model. We use a linear model:

$$F(s) = \sum_k \alpha_k \phi_k(s) \quad (5)$$

where α_k are the model parameters, and $\phi_k(.)$ are the feature functions from Sections 3.1–3.3, e.g. confidence score, similarity to L, and score for the utility of translation units. Using 20K of Spanish unlabeled text we compared the r^2 correlation coefficient between each of these scores which, apart from the arithmetic and geometric versions of the same score, showed low correlation. And so the information they provide should be complementary to each other.

We train the parameters in (5) using two bilingual development sets dev1 and dev2, the sentences in dev1 can be ranked with respect to the amount by which each particular sentence improves the BLEU score of the retrained[3] SMT model on dev2. Having this ranking, we look for the weight vector which produces the same ordering of sentences. As an alternative to this method (or its computationally demanding generalization in which instead of a single sentence, several sets of sentences of size k are selected and ranked) we use a hill climbing search on the surface of dev2's BLEU score. For a fixed value of the weight vector, dev1 sentences are ranked and then the top-k output is selected and the amount of improvement the retrained SMT system gives on dev2's BLEU score is measured. Starting from a random initial value for α_k's, we improve one dimension at a time and traverse the discrete grid placed on the values of the weight vector. Starting with a coarse grid, we make it finer when we get stuck in local optima during hill climbing.

3.5 Hierarchical Adaptive Sampling (HAS)

(Dasgupta and Hsu, 2008) propose a technique for sample selection that, under certain settings, is guaranteed to be no worse than random sampling. Their method exploits the cluster structure (if there is any) in the unlabeled data. Ideally, querying the label of only one of the data points in a cluster would be enough to determine the label of the other data points in that cluster. Their method requires that the data set is provided in the form of a tree representing a hierarchical clustering of the data. In AL for SMT, such a unique clustering of the unlabeled data would be inappropriate or ad-hoc. For this reason, we present a new algorithm inspired by the rationale provided in (Dasgupta and Hsu, 2008) that can be used in our setting, where we construct a tree-based partition of the data dynamically[4]. This dynamic tree construction allows us to extend the HAS algorithm from classifiers to the SMT task.

The algorithm adaptively samples sentences from U while building a hierarchical clustering of the sentences in U (see Fig. 1 and Algorithm 2). At any iteration, first we retrain the SMT model and translate all monolingual sentences. At this point one monolingual set of sentences represented by one of the tree leaves is chosen for further partitioning: a leaf H is chosen which has the lowest average decoder confidence score for its sentence translations. We then rank all sentences in H based on their similarity to L and put the top $\alpha|H|$ sentences in H_1 and the rest in H_2. To select K sentences, we randomly sample βK sentences from H_1 and $(1-\beta)K$ sentences from H_2 and ask a human for their translations.

3.6 Reverse Model (Reverse)

While a translation system $M_{F\to E}$ is built from language F to language E, we also build a translation system in the reverse direction $M_{E\to F}$. To measure how informative a monolingual sentence $\mathbf{f}$ is, we translate it to English by $M_{F\to E}$ and then project

[2]To see how different rankings can be combined, see (Reichart et al., 2008) which proposes this for multi-task AL.

[3]Here the retrained SMT model is the one learned by adding a particular sentence from dev1 into L.

[4]The dynamic nature of the hierarchy comes from two factors: (1) selecting a leaf node for splitting, and (2) splitting a leaf node based on its similarity to the growing L.

Algorithm 2 Hierarchical-Adaptive-Sampling

$M_{F\to E} = \textbf{train}(L, \emptyset)$
Initialize the tree T by setting its root to U
$v :=$ root(T)
for $t = 1, 2, ...$ **do**
  // rank and split sentence in v
  $X_1, X_2 :=$ Partition(L, v, α)
  // randomly sample and remove sents from X_i
  $Y_1, Y_2 :=$ Sampling(X_1, X_2, β)
  // make X_i children of node v in the tree T
  $T :=$ UpdateTree(X_1, X_2, v, T)
  // Y_i^+ has sents in Y_i together with human trans
  $L := L \cup Y_1^+ \cup Y_2^+$
  $M_{F\to E} = \textbf{train}(L, U)$
  for all leaves $l \in T$ **do**
    $Z[l] :=$ Average normalized confidence scores of sentence translations in l
  end for
  $v :=$ BestLeaf(T, Z)
  Monitor the performance on the test set
end for

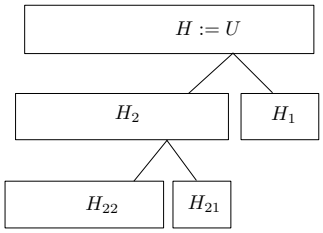

Figure 1: Adaptively sampling the sentences while constructing a hierarchical clustering of U.

the translation back to French using $M_{E\to F}$. Denote this *reconstructed* version of the original French sentence by $\tilde{\mathbf{f}}$. Comparing $\mathbf{f}$ with $\tilde{\mathbf{f}}$ using BLEU (or other measures) can tell us how much information has been lost due to our direct and/or reverse translation systems. The sentences with higher information loss are selected for translation by a human.

4 Experiments

The SMT system we applied in our experiments is PORTAGE (Ueffing et al., 2007a). The models (or features) which are employed by the decoder are: (a) one or several phrase table(s), which model the translation direction $p(\mathbf{f} \mid \mathbf{e})$, (b) one or several n-gram language model(s) trained with the SRILM toolkit (Stolcke, 2002); in the experiments reported here, we used 4-gram models on the NIST data, and a trigram model on EuroParl, (c) a distortion model which assigns a penalty based on the number of source words which are skipped when generating a new target phrase, and (d) a word penalty. These different models are combined log-linearly. Their weights are optimized w.r.t. BLEU score using the algorithm described in (Och, 2003). This is done on a development corpus which we will call dev1 in this paper.

corpus	language	use	sentences
EuroParl	Fr,Ge,Sp	in-dom L	5K
		in-dom U	20K
		in-dom dev	2K
		in-dom test	2K
See Sec. 4.2	Bangla	in-dom L	11K
		in-dom U	20K
		in-dom dev	450
		in-dom test	1K
Hansards	Fr	out-dom L	5K

Table 1: Specification of different data sets we will use in experiments. The target language is English in the bilingual sets, and the source languages are either French (Fr), German (Ge), Spanish (Sp), or Bangla.

The weight vectors in n-gram and similarity methods are set to $(.15, .2, .3, .35)$ to emphasize longer n-grams. We set $\alpha = \beta = .35$ for HAS, and use the 100-best list of translations when identifying candidate phrases while setting the maximum phrase length to 10. We set $\epsilon = .5$ to smooth probabilities when computing scores based on translation units.

4.1 Simulated Low Density Language Pairs

We use three language pairs (French-English, German-English, Spanish-English) to compare all of the proposed sentence selection strategies in a simulated AL setting. The training data comes from EuroParl corpus as distributed for the shared task in the NAACL 2006 workshop on statistical machine translation (WSMT06). For each language pair, the first 5K sentences from its bilingual corpus constitute L, and the next 20K sentences serve as U where the target side translation is ignored. The size of L was taken to be 5K in order to be close to a realistic setting in SMT. We use the first 2K sentences from the test sets provided for WSMT06, which are in-domain, as our test sets. The corpus statistics are summarized in Table 1. The results are shown in Fig. 2. After building the initial MT systems, we se-

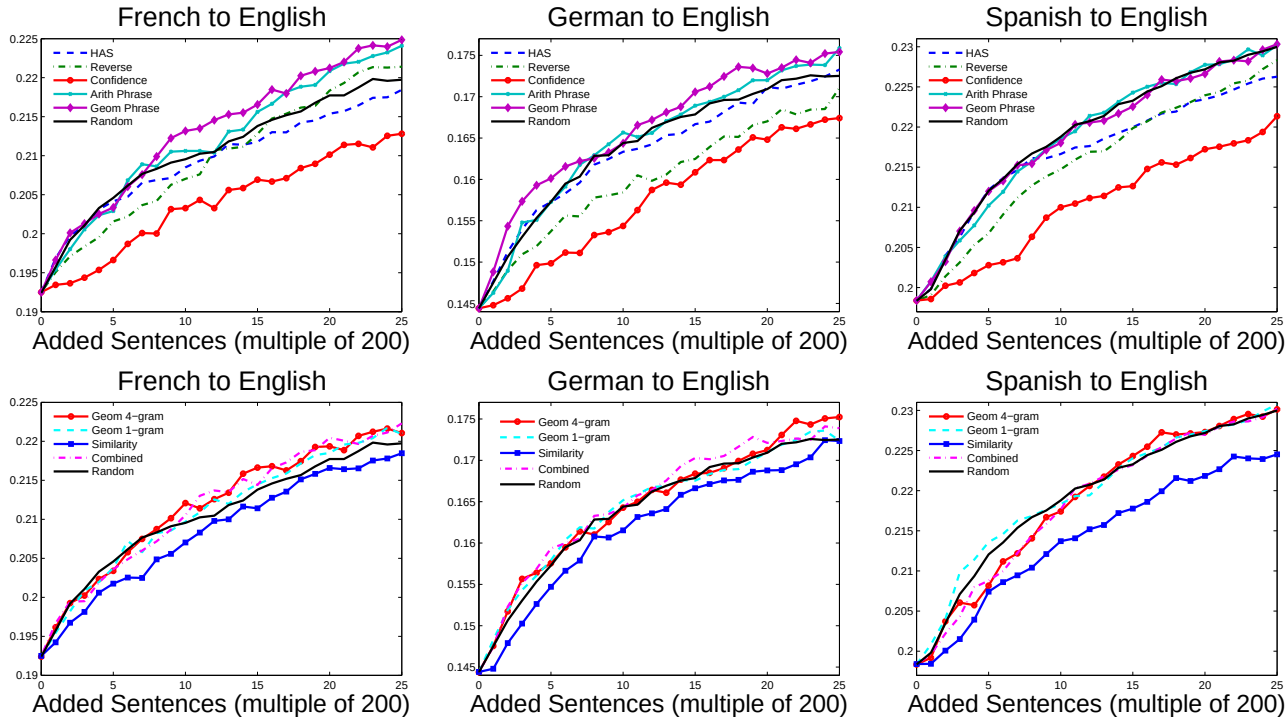

Figure 2: BLEU scores for different sentence selection strategies per iteration of the AL algorithm. Plots at the top show the performance of sentence selection methods which depend on the target language in addition to the source language (hierarchical adaptive sampling, reverse model, decoder confidence, average and geometric phrase-based score), and plots at the bottom show methods which are independent of the target language (geometric 4-gram and 1-gram, similarity to L, and random sentence selection baseline).

lect and remove 200 sentence from U in each iteration and add them together with translations to L for 25 iterations. Each experiment which involves randomness, such as random sentence selection baseline and HAS, is averaged over three independent runs. Selecting sentences based on the phrase-based utility score outperforms the strong random sentence selection baseline and other methods (Table 2). Decoder confidence performs poorly as a criterion for sentence selection in this setting, and HAS which is built on top of confidence and similarity scores outperforms both of them. Although choosing sentences based on their n-gram score ignores the relationship between source and target languages, this methods outperforms random sentence selection.

4.2 Realistic Low Density Language Pair

We apply active learning to the Bangla-English machine translation task. Bangla is the official language of Bangladesh and second most spoken language in India. It has more than 200 million speakers around the world. However, Bangla has few available language resources, and lacks resources for machine translation. In our experiments, we use training data provided by the Linguistic Data Consortium[5] containing $\sim$11k sentences. It contains newswire text from the BBC Asian Network and some other South Asian news websites. A bilingual Bangla-English dictionary collected from different websites was also used as part of the training set which contains around 85k words. Our monolingual corpus[6] is built by collecting text from the *Prothom Alo* newspaper, and contains all the news available for the year of 2005 – including magazines and periodicals. The corpus has 18,067,470 word tokens and 386,639 word types. For our language model we used data from the English section of EuroParl. The

[5]LDC Catalog No.: LDC2008E29.

[6]Provided by the Center for Research on Bangla Language Processing, BRAC University, Bangladesh.

development set used to optimize the model weights in the decoder, and test set used for evaluation was taken from the same LDC corpus mentioned above.

We applied our active learning framework to the problem of creating a larger Bangla-English parallel text resource. The second author is a native speaker of Bangla and participated in the active learning loop, translating 100 sentences in each iteration. We compared a smaller number of alternative methods to keep the annotation cost down. The results are shown in Fig. 3. Unlike the simulated setting, in this realistic setting for AL, adding more human translation does not always result in better translation performance[7]. Geom 4-gram and Geom phrase are the features that prove most useful in extracting useful sentences for the human expert to translate.

4.3 Domain Adaptation

In this section, we investigate the behavior of the proposed methods when unlabeled data U and test data T are in-domain and parallel training text L is out-of-domain.

We report experiments for French to English translation task where T and development sets are the same as those in section 4.1 but the bilingual training data come from Hansards[8] corpus. The domain is similar to EuroParl, but the vocabulary is very different. The results are shown in Fig. 4, and summarized in Table 3. As expected, unigram based sentence selection performs well in this scenario since it quickly expands the lexicon set of the bilingual data in an effective manner (Fig 5). By ignoring sentences for which the translations are already known based on L, it does not waste resources. On the other hand, it raises the importance of high frequency words in U. Interestingly, decoder confidence is also a good criterion for sentence selection in this particular case.

[7]This is likely due to the fact that the translator in the AL loop was not the same as the original translator for the labeled data.

[8]The transcription of official records of the Canadian Parliament as distributed at http://www.isi.edu/natural-language/download/hansard

Lang. Pair	Geom Phrase bleu%	per%	wer%	Random (baseline) bleu%	per%	wer%
Fr-En	22.49	27.99	38.45	21.97	28.31	38.80
Gr-En	17.54	31.51	44.28	17.25	31.63	44.41
Sp-En	23.03	28.86	39.17	23.00	28.97	39.21

Table 2: Phrase-based utility selection is compared with random sentence selection baseline with respect to BLEU, wer (word *error* rate), and per (position independent word *error* rate) across three language pairs.

method	bleu%	per%	wer%
Geom 1-gram	**14.92**	34.83	46.06
Confidence	**14.74**	35.02	46.11
Random (baseline)	14.11	35.28	46.47

Table 3: Comparison of methods in domain adaptation scenario. The bold numbers show statistically significant improvement with respect to the baseline.

5 Related Work

Despite the promise of active learning for SMT for domain adaptation and low-density/low-resource languages, there has been very little work published on this issue. 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. (Mohit and Hwa, 2007) provide a technique to classify phrases as difficult to translate (DTP), and incorporate human translations for these phrases. Their approach is different from AL: they use human translations for DTPs in order to improve translation output in the decoder. There is work on sampling sentence pairs for SMT (Kauchak, 2006; Eck et al., 2005) but the goal

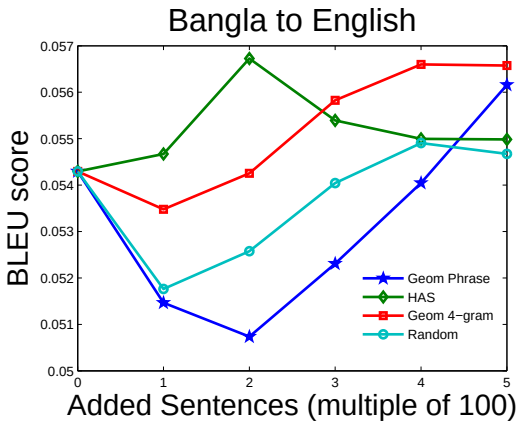

Figure 3: Improving Bangla to English translation performance using active learning.

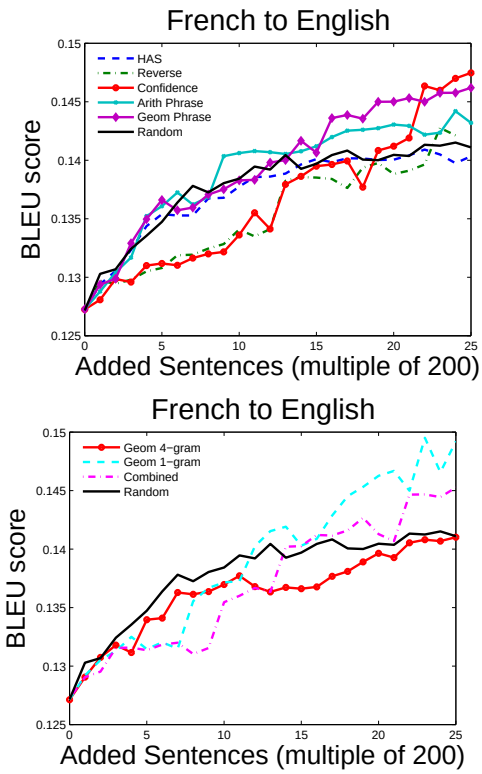

Figure 4: Performance of different sentence selection methods for domain adaptation scenario.

	Fr2En	Ge2En	Sp2En	Ha2En
Avg # of trans	1.30	1.26	1.27	1.30
	1.24	1.25	1.20	1.26
	1.22	1.23	1.19	1.24
	1.22	1.24	1.19	1.24
Avg phrase len	2.85	2.56	2.85	2.85
	3.47	2.74	3.54	3.17
	3.95	3.34	3.94	3.48
	3.58	2.94	3.63	3.36
# of phrases	27,566	29,297	30,750	27,566
	78,026	64,694	93,593	108,787
	79,343	63,191	93,276	115,177
	77,394	65,198	94,597	115,671
# unique events	31,824	33,141	34,937	31,824
	103,124	84,512	125,094	117,214
	86,210	69,357	100,176	127,314
	84,787	72,280	101,636	128,912

Table 4: Average number of english phrases per source language phrase, average length of the source language phrases, number of source language phrases, and number of phrase pairs which has been seen once in the phrase tables across three language pairs (French text taken from Hansard is abbreviated by 'Ha'). From top to bottom in each row, the numbers belong to: before starting AL, and after finishing AL based on 'Geom Phrase', 'Confidence', and 'Random'.

has been to limit the amount of training data in order to reduce the memory footprint of the SMT decoder. To compute this score, (Eck et al., 2005) use n-gram features very different from the n-gram features proposed in this paper. (Kato and Barnard, 2007) implement an AL system for SMT for language pairs with limited resources (En-Xhosa, En-Zulu, En-Setswana and En-Afrikaans), but the experiments are on a very small simulated data set. The only feature used is the confidence score of the SMT system, which we showed in our experiments is not a reliable feature.

6 Conclusions

We provided a novel active learning framework for SMT which utilizes both labeled and unlabeled data. Several sentence selection strategies were proposed and comprehensively compared across three simulated language pairs and a realistic setting of Bangla-English translation with scarce resources. Based on our experiments, we conclude that paying attention to units of translations, i.e. words and candidate phrases in particular, is essential to sentence selection in AL-SMT. Increasing the coverage of the bilingual training data is important but is not the only factor (see Table 4 and Fig. 5). For example, decoder confidence for sentence selection has low coverage (in terms of new words), but performs well in the domain adaptation scenario and performs poorly otherwise. In future work, we plan to explore selection methods based on potential phrases, adaptive sampling using features other than decoder confidence and the use of features from confidence estimation in MT (Ueffing and Ney, 2007).

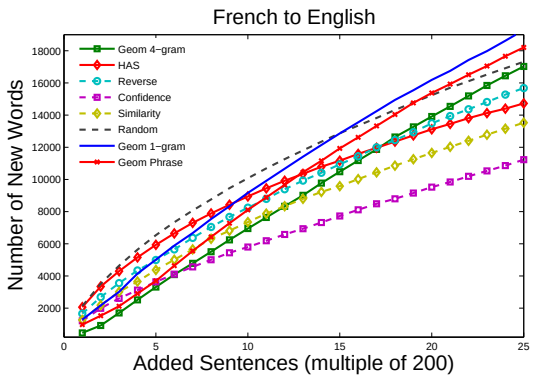

Figure 5: Number of words in domain adaptation scenario.

References

Chris Callison-burch. 2003. Active learning for statistical machine translation. In *PhD Proposal, Edinburgh University*.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. In *Machine Learning Journal*.

Sanjoy Dasgupta and Daniel Hsu. 2008. Hierarchical sampling for active learning. In *proceedings of International Conference on Machine Learning*.

Matthias Eck, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based in n-gram frequency and tf-idf. In *proceedings of International Workshop on Spoken Language Translation (IWSLT)*.

R.S.M. Kato and E. Barnard. 2007. Statistical translation with scarce resources: a south african case study. *SAIEE Africa Research Journal*, 98(4):136–140, December.

David Kauchak. 2006. Contribution to research on machine translation. In *PhD Thesis, University of California at San Diego*.

Behrang Mohit and Rebecca Hwa. 2007. Localization of difficult-to-translate phrases. In *proceedings of the 2nd ACL Workshop on Statistical Machine Translations*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *proceedings of International Conference on Spoken Language Processing (ICSLP)*.

Marco Turchi, Tijl De Bie, and Nello Cristianini. 2008. Learning performance of a machine translation system: a statistical and computational analysis. In *proceedings of the Third Workshop on Statistical Machine Translation*. Association for Computational Linguistics (ACL).

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Computational Linguistics*, 33(1):9–40.

N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson. 2007a. NRC's Portage system for WMT 2007. In *Proc. ACL Workshop on SMT*.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007b. Transductive learning for statistical machine translation. In *proceedings of Annual Meeting of the Association for Computational Linguistics (ACL)*.