Discriminative Sample Selection for Statistical Machine Translation*

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA, U.S.A.
{sanantha,rprasad,stallard,prem}@bbn.com

Abstract

Production of parallel training corpora for the development of statistical machine translation (SMT) systems for resource-poor languages usually requires extensive manual effort. Active sample selection aims to reduce the labor, time, and expense incurred in producing such resources, attaining a given performance benchmark with the smallest possible training corpus by choosing informative, non-redundant source sentences from an available candidate pool for manual translation. We present a novel, discriminative sample selection strategy that preferentially selects batches of candidate sentences with constructs that lead to erroneous translations on a held-out development set. The proposed strategy supports a built-in diversity mechanism that reduces redundancy in the selected batches. Simulation experiments on English-to-Pashto and Spanish-to-English translation tasks demonstrate the superiority of the proposed approach to a number of competing techniques, such as random selection, dissimilarity-based selection, as well as a recently proposed semi-supervised active learning strategy.

1 Introduction

Resource-poor language pairs present a significant challenge to the development of statistical machine translation (SMT) systems due to the latter's dependence on large parallel texts for training. Bilingual human experts capable of producing the requisite data resources are often in short supply, and the task of preparing high-quality parallel corpora is laborious and expensive. In light of these constraints, an attractive strategy is to construct the smallest possible parallel training corpus with which a desired performance benchmark may be achieved.

Such a corpus may be constructed by selecting the most informative instances from a large collection of source sentences for translation by a human expert, a technique often referred to as *active learning*. A SMT system trained with sentence pairs thus generated is expected to perform significantly better than if the source sentences were chosen using, say, a naïve random sampling strategy.

Previously, Eck et al. (2005) described a selection strategy that attempts to maximize coverage by choosing sentences with the highest proportion of previously unseen n-grams. Depending on the composition of the candidate pool with respect to the domain, this strategy may select irrelevant outliers. They also described a technique based on TF-IDF to de-emphasize sentences similar to those that have already been selected, thereby encouraging diversity. However, this strategy is bootstrapped by random initial choices that do not necessarily favor sentences that are difficult to translate. Finally, they worked exclusively with the source language and did not use any SMT-derived features to guide selection.

Haffari et al. (2009) proposed a number of features, such as similarity to the seed corpus, translation probability, n-gram and phrase coverage, etc., that drive data selection. They also proposed a model in which these features combine linearly to predict a rank for each candidate sentence. The

*Distribution Statement "A" (Approved for Public Release, Distribution Unlimited)

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 626–635,
MIT, Massachusetts, USA, 9-11 October 2010.

top-ranked sentences are chosen for manual translation. However, this approach requires that the pool have the same distributional characteristics as the development sets used to train the ranking model. Additionally, batches are chosen atomically. Since similar or identical sentences in the pool will typically meet the selection criteria simultaneously, this can have the undesired effect of choosing redundant batches with low diversity.

The semi-supervised active learning strategy proposed by Ananthakrishnan et al. (2010) uses multilayer perceptrons (MLPs) to rank candidate sentences based on various features, including domain representativeness, translation difficulty, and batch diversity. A greedy, incremental batch construction technique encourages diversity. While this strategy was shown to be superior to random as well as n-gram based dissimilarity selection, its coarse granularity (reducing a candidate sentence to a low-dimensional feature vector for ranking) makes it unsuitable for many situations. In particular, it is seen to have little or no benefit over random selection when there is no logical separation of the candidate pool into "in-domain" and "out-of-domain" subsets.

This paper introduces a novel, active sample selection technique that identifies translation errors on a held-out development set, and preferentially selects candidate sentences with constructs that are incorrectly translated in the former. A discriminative pairwise comparator function, trained on the ranked development set, is used to order candidate sentences and pick sentences that provide maximum potential reduction in translation error. The feature functions that power the comparator are updated after each selection to encourage batch diversity. In the following sections, we provide details of the proposed sample selection approach, and describe simulation experiments that demonstrate its superiority over a number of competing strategies.

2 Error-Driven Active Learning

Traditionally, unsupervised selection strategies have dominated the active learning literature for natural language processing (Hwa, 2004; Tang et al., 2002; Shen et al., 2004). Sample selection for SMT has followed a similar trend. The work of Eck et al. (2005) and most of the techniques proposed by Haffari et al. (2009) fall in this category. Notable exceptions include the linear ranking model of Haffari et al. (2009) and the semi-supervised selection technique of Ananthakrishnan et al. (2010), both of which use one or more held-out development sets to train and tune the sample selector. However, while the former uses the posterior translation probability and the latter, a sentence-level confidence score as part of the overall selection strategy, current active learning techniques for SMT do not explicitly target the sources of error.

Error-driven active learning attempts to choose candidate instances that potentially maximize error reduction on a reference set (Cohn et al., 1996; Meng and Lee, 2008). In the context of SMT, this involves decoding a held-out development set with an existing baseline (seed) SMT system. The selection algorithm is then trained to choose, from the candidate pool, sentences containing constructs that give rise to translation errors on this set. Assuming perfect reference translations and word alignment in subsequent SMT training, these sentences provide maximum potential reduction in translation error with respect to the seed SMT system. It is a *supervised* approach to sample selection. We assume the following are available.

- A seed parallel corpus $\mathbf{S}$ for training the initial SMT system.
- A candidate pool of monolingual source sentences $\mathbf{P}$ from which samples must be selected.
- A held-out development set $\mathbf{D}$ for training the selection algorithm and for tuning the SMT.
- A test set $\mathbf{T}$ for evaluating SMT performance.

We further make the following reasonable assumptions: (a) the development set $\mathbf{D}$ and the test set $\mathbf{T}$ are drawn from the same distribution and (b) the candidate pool $\mathbf{P}$ consists of both in- and out-of-domain source sentences, as well as an allowable level of redundancy (similar or identical sentences).

Using translation errors on the development set to drive sample selection has the following advantages over previously proposed active learning strategies for SMT.

- The seed training corpus $\mathbf{S}$ need not be derived from the same distribution as $\mathbf{D}$ and $\mathbf{T}$. The seed SMT system can be trained with any available

parallel corpus for the specified language pair. This is very useful if, as is often the case, little or no in-domain training data is available to bootstrap the SMT system. This removes a critical restriction present in the semi-supervised approach of Ananthakrishnan et al. (2010).

- Sentences chosen are guaranteed to be relevant to the domain, because selection is based on n-grams derived from the development set. This alleviates potential problems with approaches suggested by Eck et al. (2005) and several techniques used by Haffari et al. (2009), where irrelevant outliers may be chosen simply because they contain previously unseen n-grams, or are deemed difficult to translate.
- The proposed technique seeks to minimize held-out translation error rather than maximize training-set coverage. This is the more intuitive, direct approach to sample selection for SMT.
- Diversity can be encouraged by preventing n-grams that appear in previously selected sentences from playing a role in choosing subsequent sentences. This provides an efficient alternative to the cumbersome "batch diversity" feature proposed by Ananthakrishnan et al. (2010).

The proposed implementation of error-driven active learning for SMT, discriminative sample selection, is described in the following section.

3 Discriminative Sample Selection

The goal of active sample selection is to induce an ordering of the candidate instances that satisfies an objective criterion. Eck et al. (2005) ordered candidate sentences based on the frequency of unseen n-grams. Haffari et al. (2009) induced a ranking based on unseen n-grams, translation difficulty, etc., as well as one that attempted to incrementally maximize BLEU using two held-out development sets. Ananthakrishnan et al. (2010) attempted to order the candidate pool to incrementally maximize source n-gram coverage on a held-out development set, subject to difficulty and diversity constraints.

In the case of error-driven active learning, we attempt to learn an ordering model based on errors observed on the held-out development set $\mathbf{D}$. We achieve this in an innovative fashion by casting the ranking problem as a pairwise sentence comparison problem. This approach, inspired by Ailon and Mohri (2008), involves the construction of a binary classifier functioning as a relational operator that can be used to order the candidate sentences. The pairwise comparator is trained on an ordering of $\mathbf{D}$ that ranks constituent sentences in decreasing order of the number of translation errors. The comparator is then used to rank the candidate pool in decreasing order of potential translation error reduction.

3.1 Maximum-Entropy Pairwise Comparator

Given a pair of source sentences (u, v), we define, adopting the notation of Ailon and Mohri (2008), the pairwise comparator $h(u, v)$ as follows:

$$h(u,v) = \begin{cases} 1, & u < v \\ 0, & u >= v \end{cases} \quad (1)$$

In Equation 1, the binary comparator $h(u, v)$ plays the role of the "less than" ("$<$") relational operator, returning 1 if u is preferred to v in an ordered list, and 0 otherwise. As detailed in Ailon and Mohri (2008), the comparator must satisfy the constraint that $h(u, v)$ and $h(v, u)$ be complementary, i.e. $h(u, v) + h(v, u) = 1$ to avoid ambiguity. However, it need not satisfy the triangle inequality.

We implement $h(u, v)$ as a combination of discriminative maximum entropy classifiers triggered by feature functions drawn from n-grams of u and v. We define $p(u, v)$ as the conditional posterior probability of the Bernoulli event $u < v$ given (u, v) as shown in Equation 2.

$$p(u,v) = Pr(u < v \mid u, v) \quad (2)$$

In our implementation, $p(u, v)$ is the output of a binary maximum-entropy classifier trained on the development set. However, this implementation poses two problems.

First, if we use constituent n-grams of u and v as feature functions to trigger the classifier, there is no way to distinguish between (u, v) and (v, u) as they will trigger the same feature functions. This will result in identical values for $p(u, v)$ and $p(v, u)$, a contradiction. We resolve this issue by introducing a set of "complementary" feature functions, which are formed by simply appending a recognizable identifier to the existing n-gram feature func-

u: how are you
v: i am going

$\mathbf{f}(u) = \{$how:1, are:1, you:1, how*are:2, are*you:2, how*are*you:3$\}$
$\mathbf{f}(v) = \{$i:1, am:1, going:1, i*am:2, am*going:2, i*am*going:3$\}$

$\mathbf{f}'(u) = \{$!how:1, !are:1, !you:1, !how*are:2, !are*you:2, !how*are*you:3$\}$
$\mathbf{f}'(v) = \{$!i:1, !am:1, !going:1, !i*am:2, !am*going:2, !i*am*going:3$\}$

Table 1: Standard and complementary trigram feature functions for a source pair (u, v).

tions. Then, to evaluate $p(u, v)$, for instance, we invoke the classifier with standard feature functions for u and complementary feature functions for v. Similarly, $p(v, u)$ is evaluated by triggering complementary feature functions for u and standard feature functions for v. Table 1 illustrates this with a simple example.

Note that each feature function is associated with a real value, whose magnitude is an indicator of its importance. In our implementation, an n-gram feature function (standard or complementary) receives a value equal to its length. This is based on our intuition that longer n-grams play a more important role in dictating SMT performance.

Second, the introduction of complementary triggers implies that evaluation of $p(u, v)$ and $p(v, u)$ now involves disjoint sets of feature functions. Thus, $p(u, v)$ is not guaranteed to satisfy the complementarity condition imposed on $h(u, v)$, and therefore cannot directly be used as the binary pairwise comparator. We resolve this by normalizing across the two possible permutations, as follows:

$$h'(u, v) = \frac{p(u, v)}{p(u, v) + p(v, u)} \quad (3)$$

$$h'(v, u) = \frac{p(v, u)}{p(u, v) + p(v, u)} \quad (4)$$

Since $h'(u, v) + h'(v, u) = 1$, the complementarity constraint is now satisfied, and $h(u, v)$ is just a binarized (thresholded) version of $h'(u, v)$. Thus, the binary pairwise comparator can be constructed from the permuted classifier outputs.

3.2 Training the Pairwise Comparator

Training the maximum-entropy classifier for the pairwise comparator requires a set of target labels and input feature functions, both of which are derived from the held-out development set $\mathbf{D}$. We begin by decoding the source sentences in $\mathbf{D}$ with the seed SMT system, followed by error analysis using the Translation Edit Rate (TER) measure (Snover et al., 2006). TER measures translation quality by computing the number of edits (insertions, substitutions, and deletions) and shifts required to transform a translation hypothesis to its corresponding reference. We then rank $\mathbf{D}$ in decreasing order of the number of post-shift edits, i.e. the number of insertions, substitutions, and deletions after the shift operation is completed. Since shifts are often due to word re-ordering issues within the SMT decoder (especially for phrase-based systems), we do not consider them as errors for the purpose of ranking $\mathbf{D}$. Sentences at the top of the ordered list $\mathbf{D}'$ contain the maximum number of translation errors.

For each pair of sentences $(u, v) : u < v$ in $\mathbf{D}'$, we generate two training entries. The first, signifying that u appears before v in $\mathbf{D}'$, assigns the label *true* to a trigger list consisting of standard feature functions derived from u, and complementary feature functions derived from v. The second, reinforcing this observation, assigns the label *false* to a trigger list consisting of complementary feature functions from u, and standard feature functions from v. The labeled training set (feature:label pairs) for the comparator can be expressed as follows:

$$\begin{aligned} \forall(u, v) \quad \in \quad & \mathbf{D}' : u < v, \\ & \{\mathbf{f}(u) \quad \mathbf{f}'(v)\} : true \\ & \{\mathbf{f}'(u) \quad \mathbf{f}(v)\} : false \end{aligned}$$

Thus, if there are d sentences in $\mathbf{D}'$, we obtain a total of $d(d-1)$ labeled examples to train the comparator. We use the standard L-BFGS optimization

algorithm (Liu and Nocedal, 1989) to estimate the parameters of the maximum entropy model.

3.3 Greedy Discriminative Selection

The discriminatively-trained pairwise comparator can be used as a relational operator to sort the candidate pool $\mathbf{P}$ in decreasing order of potential translation error reduction. A batch of pre-determined size K can then be selected from the top of this list to augment the existing SMT training corpus. Assuming the pool contains N candidate sentences, and given a fast sorting algorithm such as Quicksort, the complexity of this strategy is $O(N \log N)$. Batches can be selected iteratively until a specified performance threshold is achieved.

A potential downside of this approach reveals itself when there is redundancy in the candidate pool. Since the batch is selected in a single atomic operation from the sorted candidates, and because similar or identical sentences will typically occupy the same range in the ordered list, it is likely that this approach will result in batches with low diversity. Whereas we desire diverse batches for better coverage and efficient use of manual translation resources. This issue was previously addressed in Shen et al. (2004) in the context of named-entity recognition, where they used a two-step procedure to first select the most informative and representative samples, followed by a diversity filter. Ananthakrishnan et al. (2010) used a greedy, incremental batch construction strategy with an integrated, explicit batch diversity feature as part of the ranking model. Based on these ideas, we design a greedy selection strategy using the discriminative relational operator.

Rather than perform a full sort on $\mathbf{P}$, we simply invoke the $min_{h(u,v)}(\cdots)$ function to find the sentence that potentially minimizes translation error. The subscript indicates that our implementation of this function utilizes the discriminative relational operator trained on the development set $\mathbf{D}$. The best choice sentence s is then added to our batch at the current position (we begin with an empty batch). We then remove the standard and complementary feature functions $\mathbf{f}(s)$ and $\mathbf{f}'(s)$ triggered by s from the global pool of feature functions obtained from $\mathbf{D}$, so that they do not play a role in the selection of subsequent sentences for the batch. Subsequently, a candidate sentence that is similar or identical to s will not be preferred, because the feature functions that previously caused it to rank highly will no longer trigger. Algorithm 1 summarizes our selection strategy in pseudocode. Since each call to $min_{h(u,v)}(\cdots)$ is $O(N)$, the overall complexity of greedy discriminative selection is $O(K \cdot N)$.

Algorithm 1 Greedy Discriminative Selection

$\mathbf{B} \leftarrow ()$
for $k = 1$ to K **do**
  $s \leftarrow min_{h(u,v)}(\mathbf{P})$
  $B(k) \leftarrow s$
  $\mathbf{P} \leftarrow \mathbf{P} - \{s\}$
  $\mathbf{f}(\mathbf{D}) \leftarrow \mathbf{f}(\mathbf{D}) - \mathbf{f}(s)$
  $\mathbf{f}'(\mathbf{D}) \leftarrow \mathbf{f}'(\mathbf{D}) - \mathbf{f}'(s)$
end for
return $\mathbf{B}$

4 Experiments and Results

We conduct a variety of simulation experiments with multiple language pairs (English-Pashto and Spanish-English) and different data configurations in order to demonstrate the utility of discriminative sample selection in the context of resource-poor SMT. We also compare the performance of the proposed strategy to numerous competing active and passive selection methods as follows:

- *Random*: Source sentences are uniformly sampled from the candidate pool $\mathbf{P}$.
- *Similarity*: Choose sentences from $\mathbf{P}$ with the highest fraction of n-gram overlap with the seed corpus $\mathbf{S}$.
- *Dissimilarity*: Select sentences from $\mathbf{P}$ with the highest proportion of n-grams not seen in the seed corpus $\mathbf{S}$ (Eck et al., 2005; Haffari et al., 2009).
- *Longest*: Pick the longest sentences from the candidate pool $\mathbf{P}$.
- *Semi-supervised*: Semi-supervised active learning with greedy incremental selection (Ananthakrishnan et al., 2010).
- *Discriminative*: Choose sentences that potentially minimize translation error using a maximum-entropy pairwise comparator (proposed method).

Identical low-resource initial conditions are applied to each selection strategy so that they may be objectively compared. A very small seed corpus $\mathbf{S}$ is sampled from the available parallel training data; the remainder serves as the candidate pool. Following the literature on active learning for SMT, our simulation experiments are iterative. A fixed-size batch of source sentences is constructed from the candidate pool using one of the above selection strategies. We then look up the corresponding translations from the candidate targets (simulating an expert human translator), augment the seed corpus with the selected data, and update the SMT system with the expanded training corpus. The selected data are removed from the candidate pool. This select-update cycle is then repeated for either a fixed number of iterations or until a specified performance benchmark is attained. At each iteration, we decode the unseen test set $\mathbf{T}$ with the most current SMT configuration and evaluate translation performance in terms of BLEU as well as coverage (defined as the fraction of untranslatable source words in the target hypotheses).

We use a phrase-based SMT framework similar to Koehn et al. (2003) for all experiments.

4.1 English-Pashto Simulation

Our English-Pashto (E2P) data originates from a two-way collection of spoken dialogues, and consists of two parallel sub-corpora: a directional E2P corpus and a directional Pashto-English (P2E) corpus. Each sub-corpus has its own independent training, development, and test partitions. The directional E2P training, development, and test sets consist of 33.9k, 2.4k, and 1.1k sentence pairs, respectively. The directional P2E training set consists of 76.5k sentence pairs. The corpus was used as-is, i.e. no length-based filtering or redundancy-reduction (i.e. removal of duplicates, if any) was performed. The test-set BLEU score with the baseline E2P SMT system trained from all of the above data was 9.5%.

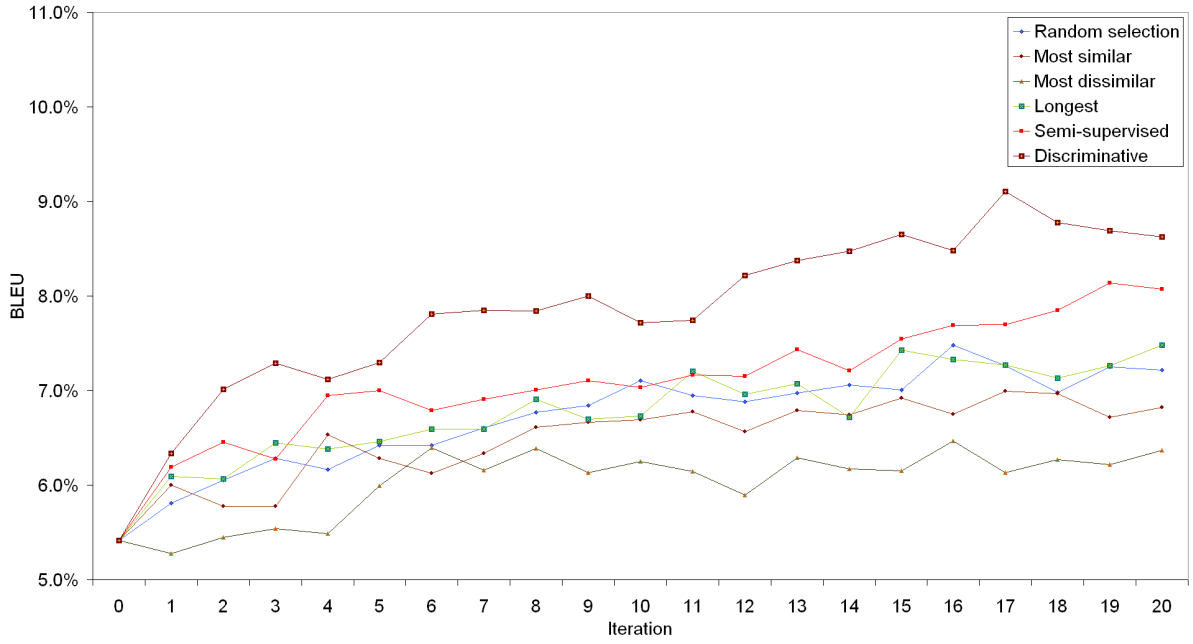
We obtained a seed training corpus by randomly sampling 1,000 sentence pairs from the directional E2P training partition. The remainder of this set, and the entire reversed P2E training partition were combined to create the pool (109.4k sentence pairs). In the past, we have observed that the reversed directional P2E data gives very little performance gain in the E2P direction even though its vocabulary is similar, and can be considered "out-of-domain" as far as the E2P translation task is concerned. Thus, our pool consists of 30% in-domain and 70% out-of-domain sentence pairs, making for a challenging active learning problem. A pool training set of 10k source sentences is sampled from this collection for the semi-supervised selection strategy, leaving us with 99.4k candidate sentences, which we use for all competing techniques. The data configuration used in this simulation is identical to Ananthakrishnan et al. (2010), allowing us to compare various strategies under the same conditions. We simulated a total of 20 iterations with batches of 200 sentences each; the original 1,000 sample seed corpus grows to 5,000 sentence pairs and the end of our simulation.

Figure 1(a) illustrates the variation in BLEU scores across iterations for each selection strategy. The proposed discriminative sample selection technique performs significantly better at every iteration than random, similarity, dissimilarity, longest, and semi-supervised active selection. At the end of 20 iterations, the BLEU score gained 3.21 points, a relative improvement of 59.3%. This was followed by semi-supervised active learning, which improved by 2.66 BLEU points, a 49.2% relative improvement. Table 2 summarizes the total number of words selected by each strategy, as well as the total area under the BLEU curve with respect to the baseline. The latter, labeled $\text{BLEU}_{\text{area}}$ and expressed in *percent-iterations*, is a better measure of the overall performance of each strategy across all iterations than comparing BLEU scores at the final iteration.

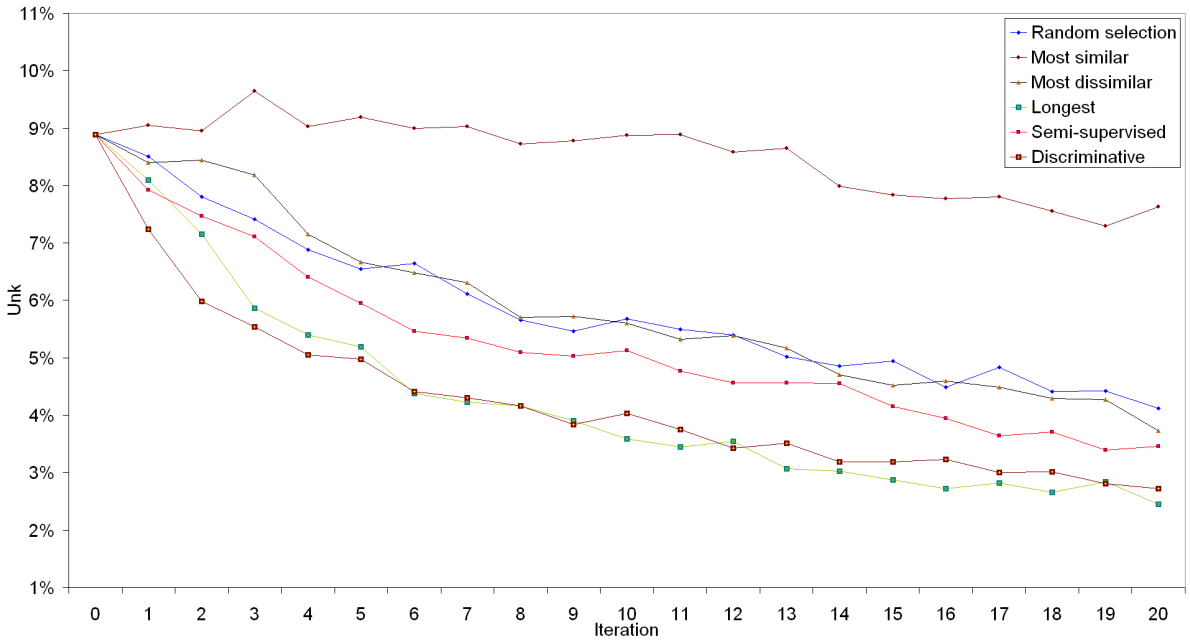
Figure 1(b) shows the variation in coverage (percentage of untranslatable source words in target hypotheses) for each selection technique. Here, discriminative sample selection was better than all other approaches except longest-sentence selection.

4.2 Spanish-English Simulation

The Spanish-English (S2E) training corpus was drawn from the Europarl collection (Koehn, 2005). To prevent length bias in selection, the corpus was filtered to only retain sentence pairs whose source ranged between 7 and 15 words (excluding punctuation). Additionally, redundancy was reduced by removing all duplicate sentence pairs. After these steps, we obtained approximately 253k sentence pairs for training. The WMT10 held-out develop-

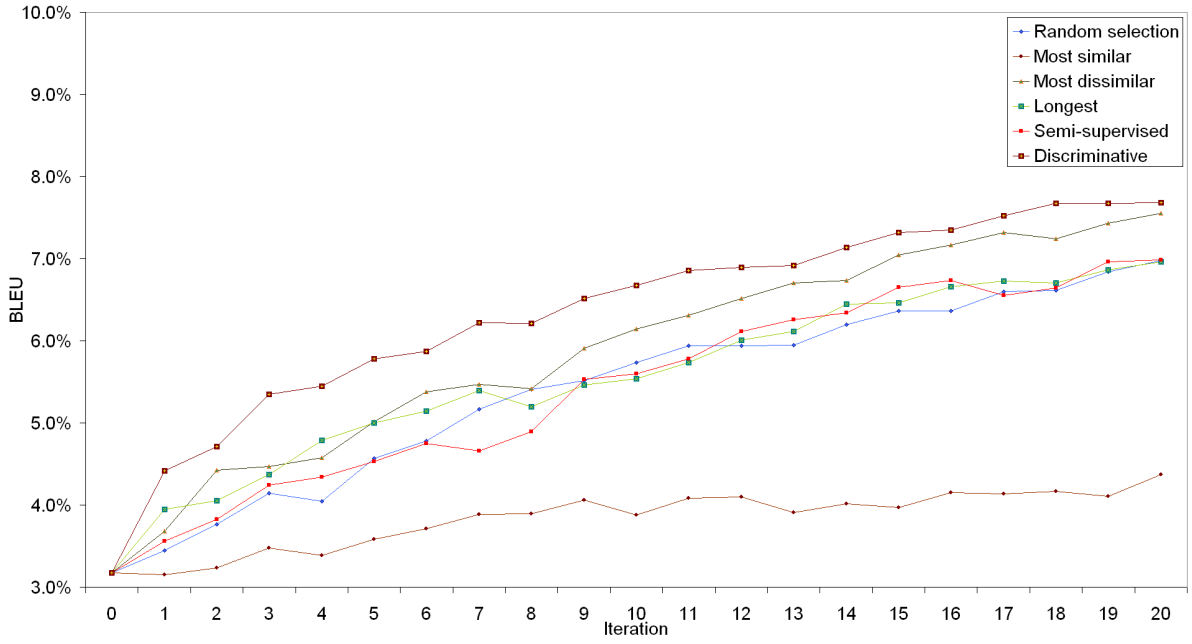


(a) Variation in BLEU (E2P)

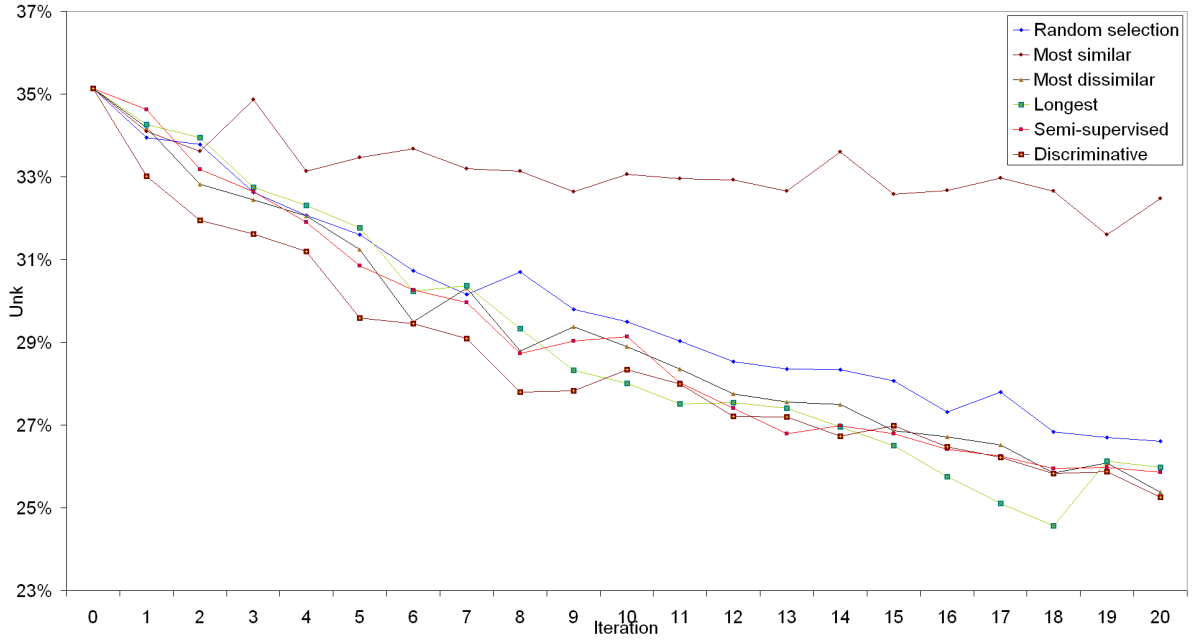


(b) Variation in coverage (E2P)

Figure 1: Simulation results for E2P data selection.



(a) Variation in BLEU (S2E)



(b) Variation in coverage (S2E)

Figure 2: Simulation results for S2E data selection.

Method	E2P size	E2P $BLEU_{area}$	S2E size	S2E $BLEU_{area}$
Random	58.1k	26.4	26.5k	45.0
Similarity	30.7k	21.9	24.7k	13.2
Dissimilarity	39.2k	12.4	24.2k	54.9
Longest	173.0k	27.5	39.6k	48.3
Semi-supervised	80.0k	34.1	27.6k	45.6
Discriminative	109.1k	49.6	31.0k	64.5

Table 2: Source corpus size (in words) and $BLEU_{area}$ after 20 sample selection iterations.

ment and test sets (2k and 2.5k sentence pairs, respectively) were used to tune our system and evaluate performance. Note that this data configuration is different from that of the E2P simulation in that there is no logical separation of the training data into "in-domain" and "out-of-domain" sets. The baseline S2E SMT system trained with all available data gave a test-set BLEU score of 17.2%.

We randomly sampled 500 sentence pairs from the S2E training partition to obtain a seed training corpus. The remainder, after setting aside another 10k source sentences for training the semi-supervised strategy, serves as the candidate pool. We again simulated a total of 20 iterations, except in this case, we used batches of 100 sentences in an attempt to obtain smoother performance trajectories. The training corpus grows from 500 sentence pairs to 2,500 as the simulation progresses.

Variation in BLEU scores and coverage for the S2E simulation are illustrated in Figures 2(a) and 2(b), respectively. Discriminative sample selection outperformed all other selection techniques across all iterations of the simulation. After 20 iterations, we obtained a 4.51 point gain in BLEU, a relative improvement of 142.3%. The closest competitor was dissimilarity-based selection, which improved by 4.38 BLEU points, a 138.1% relative improvement. The proposed method also outperformed other selection strategies in improving coverage, with significantly better results especially in the early iterations. Table 2 summarizes the number of words chosen, and $BLEU_{area}$, for each strategy.

5 Conclusion and Future Directions

Building SMT systems for resource-poor language pairs requires significant investment of labor, time, and money for the development of parallel training corpora. We proposed a novel, discriminative sample selection strategy that can help lower these costs by choosing batches of source sentences from a large candidate pool. The chosen sentences, in conjunction with their manual translations, provide significantly better SMT performance than numerous competing active and passive selection techniques.

Our approach hinges on a maximum-entropy pairwise comparator that serves as a relational operator for comparing two source sentences. This allows us to rank the candidate pool in decreasing order of potential reduction in translation error with respect to an existing seed SMT system. The discriminative comparator is coupled with a greedy, incremental selection technique that discourages redundancy in the chosen batches. The proposed technique diverges from existing work on active sample selection for SMT in that it uses machine learning techniques in an attempt to explicitly reduce translation error by choosing sentences whose constituents were incorrectly translated in a held-out development set.

While the performance of competing strategies varied across language pairs and data configurations, discriminative sample selection proved consistently superior under all test conditions. It provides a powerful, flexible, data selection front-end for rapid development of SMT systems. Unlike some selection techniques, it is also platform-independent, and can be used as-is with a phrase-based, hierarchical, syntactic, or other SMT framework.

We have so far restricted our experiments to simulations, obtaining expert human translations directly from the sequestered parallel corpus. We are now actively exploring the possibility of linking the sample selection front-end to a crowd-sourcing back-end, in order to obtain "non-expert" translations using a platform such as the Amazon Mechanical Turk.

References

Nir Ailon and Mehryar Mohri. 2008. An efficient reduction of ranking to classification. In *COLT '08: Proceedings of the 21st Annual Conference on Learning Theory*, pages 87–98.

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010. A semi-supervised batch-mode active learning strategy for improved statistical machine translation. In *CoNLL '10: Proceedings of the 14th International Conference on Computational Natural Language Learning*, pages 126–134, July.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145.

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 IWSLT*, Pittsburgh, PA, October.

Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *NAACL '09: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 415–423, Morristown, NJ, USA. Association for Computational Linguistics.

Rebecca Hwa. 2004. Sample selection for statistical parsing. *Computational Linguistics*, 30:253–276.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit X: Proceedings of the 10th Machine Translation Summit*, pages 79–86.

D. C. Liu and J. Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *Math. Program.*, 45(3):503–528.

Qinggang Meng and Mark Lee. 2008. Error-driven active learning in growing radial basis function networks for early robot learning. *Neurocomputing*, 71(7-9):1449–1461.

Dan Shen, Jie Zhang, Jian Su, Guodong Zhou, and Chew-Lim Tan. 2004. Multi-criteria-based active learning for named entity recognition. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 589–596, Morristown, NJ, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings AMTA*, pages 223–231, August.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2002. Active learning for statistical natural language parsing. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 120–127, Morristown, NJ, USA. Association for Computational Linguistics.