Source Error-Projection for Sample Selection in Phrase-Based SMT for Resource-Poor Languages

Sankaranarayanan Ananthakrishnan, Shiv Vitaladevuni, Rohit Prasad, and Prem Natarajan
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02138, U.S.A.
{sanantha,svitalad,rprasad,pnataraj}@bbn.com

Abstract

The unavailability of parallel training corpora in resource-poor languages is a major bottleneck in cost-effective and rapid deployment of statistical machine translation (SMT) technology. This has spurred significant interest in active learning for SMT to select the most informative samples from a large candidate pool. This is especially challenging when irrelevant outliers dominate the pool. We propose two supervised sample selection methods, viz. greedy selection and integer linear programming (ILP), based on a novel measure of benefit derived from error analysis. These methods support the selection of diverse and high-impact, yet relevant batches of source sentences. Comparative experiments on multiple test sets across two resource-poor language pairs (English-Pashto and English-Dari) reveal that the proposed approaches achieve BLEU scores comparable to the full system using a very small fraction of all available training data (ca. 6% for E-P and 13% for E-D). We further demonstrate that the ILP method supports global constraints of significant practical value.

1 Introduction

The laborious and time-consuming nature of producing parallel training corpora for the development of high-quality SMT systems cannot be overstated. Barring a few mainstream languages, the vast majority of language pairs can be classified as "resource-poor" as far as availability of a usable SMT system is concerned. Active learning can reduce human labor, turn-around time and monetary cost of developing SMT systems with little or no loss in translation accuracy.

In its simplest form, active learning for building parallel corpora involves selecting "high-value" samples from a large monolingual corpus of source sentences (the *candidate pool*) for translation by a bilingual human expert. The notion of "value" depends on the selection method, and can be derived using unsupervised, semi-supervised, or supervised techniques. For instance, Eck et al. (2005) define high-value source sentences as those that contain a large number of previously unseen n-grams. While it aims to increase coverage of the training set, the main deficiency of this approach is its tendency to pick irrelevant outliers if the candidate pool contains data from unrelated regimes.

Haffari et al. (2009) propose a number of features, such as similarity to the seed corpus, translation probability, n-gram and phrase coverage as unsupervised measures of the value of candidate samples. Additionally, a linear combination of these features is proposed as a supervised measure of value for ranking candidate sentences. The parameters of this model are optimized on two separate held-out bilingual development sets. The disadvantage of this approach is that it relies on the candidate pool having the same distributional characteristics as the development sets used for parameter estimation. Haffari and Sarkar (2009) explore active learning in a multilingual setting (different source languages f_d to be translated to a single target language e) using disagreement between target hypotheses generated by each of the SMT systems. For single language active learn-

This paper is based upon work supported by the DARPA TRANSTAC Program. The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 819–827,
Chiang Mai, Thailand, November 8 – 13, 2011.

ing, they use OOV phrases, i.e. source n-grams without translation choices.

Ananthakrishnan et al. (2010) propose an error-driven approach that identifies translation errors on a held-out development set, and uses this to train a discriminative pairwise comparator function that preferentially selects candidate sentences with constructs that are incorrectly translated in the development set. The chosen sentences provide maximum potential reduction in translation error. The advantage of this method over other unsupervised and semi-supervised selection strategies is that it favors domain-relevant sentences that are difficult to translate. However, its granularity is low, because it considers errors only at the sentence level. Further, the diversity constraint is implemented in a non-optimal, ad-hoc manner by deleting feature functions from the pairwise classification model.

Bloodgood and Callison-Burch (2010) explored active learning to augment training data for *high resource* language pairs. They experimented with random, shortest, and longest sentence selection, as well as a technique they refer to as *Vocab-Growth*. The latter prefers a candidate sentence that contains the most frequent n-gram not seen in the labeled training data. Again, this approach is unsuitable if the pool contains a lot of irrelevant sentences. Moreover, their system is based entirely on source language statistics, and does not use any feedback from the SMT system. As one of their examples indicates, this makes it susceptible to selecting sentences containing n-grams that were already correctly translatable.

This paper introduces a novel, fine-grained, error-driven measure of value for candidate sentences obtained by translation error analysis on a domain-relevant held-out development set. Errors identified in translation hypotheses are projected back on to the corresponding source sentences through phrase derivations from the SMT decoder. This projected error is used to obtain a "benefit value" for each source n-gram that serves as a measure of its translation difficulty. Sentence selection is posed as the problem of choosing K sentences from the candidate pool that maximize the sum of the benefit values of n-grams covered by the choice. This is a generalization of the set-covering problem, known to be NP-Complete. We present two approximate solutions: (a) an efficient greedy algorithm and (b) an integer linear programming (ILP) formulation. We compare these two methods and demonstrate their superiority to numerous competing selection strategies described in the literature.

2 Translation Error Projection

The principal advantage of error-driven sample selection (Cohn et al., 1996; Meng and Lee, 2008) over traditional unsupervised or semi-supervised active learning (Hwa, 2004; Tang et al., 2002; Shen et al., 2004) is its ability to choose instances, which, when annotated, potentially maximize error reduction of the learner on a reference set.

We assume the following data configuration for error-driven sample selection for SMT. A seed parallel corpus $\mathbf{S}$ is required to bootstrap an initial translation system. However, we do not require that this corpus be drawn from the same distribution as the testing condition. This relaxed assumption is particularly useful for developing SMT systems for resource-poor language pairs for which in-domain parallel training data may not be readily available, but a (low quality) translation system may be built using data from other domains, genres or dialects. We also assume a phrase-based SMT architecture (Koehn et al., 2003).

A held-out development (tuning) set $\mathbf{D}$ is used for optimizing the parameters of the SMT system using MERT (Och, 2003), as well as for error-analysis in guiding the proposed sample selection algorithms. System performance is evaluated on a fair test set $\mathbf{T}$. We assume that the tuning set is derived from the same distribution as the test set. The selection algorithms operate on a large pool of monolingual source sentences $\mathbf{P}$ to extract high-value samples for translation by a human expert. The candidate pool may contain any mixture of relevant and irrelevant sentences, and may also possess significant redundancy.

2.1 Error Analysis

The SMT system is bootstrapped using the seed training corpus $\mathbf{S}$. The held-out set $\mathbf{D}$ is decoded by the SMT to obtain 1-best translation hypotheses. Translation edit rate (TER) analysis (Snover et al., 2006) is used to identify errors in the hypotheses by aligning them to the target references. The TER alignment identifies a set of insertions, substitutions, deletions, and shifts that is required to transform a hypothesis to its corresponding reference. Large values of TER indicate greater dis-

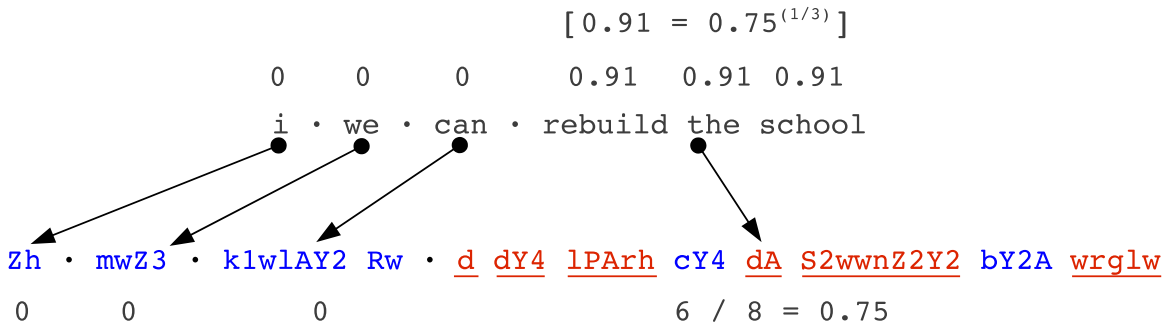


Figure 1: Example illustrating evaluation of back-projected error for a source sentence given phrase derivations and error analysis of its translation hypothesis (English-Pashto). Incorrectly hypothesized words are underlined in red.

similarity between hypotheses and references, corresponding to poor translations.

An individual target word in the hypothesis is deemed "correct" if it aligns to itself in the TER alignment. Conversely, hypothesized words corresponding to substitution or insertion errors are considered "incorrect" translations, while deletion errors are ignored. Thus, each hypothesized word can be labeled "correct" or "incorrect" based on the TER alignment, providing a fine-grained view of translation difficulty on the held-out set.

2.2 Benefit/Objective Function

TER analysis enables us to label errors at the word level for each SMT hypothesis. However, it is the knowledge of which *source words* were translated incorrectly that is useful for sample selection. Assuming there is a way to attach a label (or real value) to each source word in the held-out set $\mathbf{D}$ indicating whether it was correctly translated (or to what degree it was translated), we can compute a *benefit value* over each source n-gram. The sum of benefit values over all source n-grams can be used as an objective function that must be maximized by selecting suitable samples from the candidate pool.

The SMT decoder produces *phrase derivations* specifying the origin of each target phrase in a hypothesis, providing a convenient mechanism for approximate projection of target error labels back on to the source words. We refer to this as error back-projection. We compute, for each target phrase in the phrase derivations, the *target phrase error* as a ratio of the number of words labeled "incorrect" to the total number of words within that phrase. This quantity is then equally distributed among constituent source words using the geometric mean with respect to the number of words in the containing source phrase (obtained from the phrase derivation), to give us a back-projected error at the level of individual source words. Since the phrase derivations form a mutually exclusive partition over the source sentence, we can compute an unambiguous back-projected error value for each source word in the held-out set $\mathbf{D}$. Figure 1 illustrates this procedure with an example.

We then compute a benefit value for each source n-gram as the sum of back-projected error of the constituent source words. The set of source phrases in the SMT decoder derivations is typically a very small subset of the set of all possible n-grams of equal length. By distributing back-projected phrase error over individual target words, we are able to compute benefit values for n-grams not covered by the phrase table inventory. Finally, we sum the benefit values of source n-grams that occur multiple times in $\mathbf{D}$ to generate a table of benefit values hashed by the corresponding n-grams.

3 Sentence Selection Problem

Given a set of n-grams $\mathcal{N} = \{n_i\}_{i=1}^{m}$ from source sentences in $\mathbf{D}$ and associated benefit values $b_i \geq 0$ computed by error back-projection, the goal of sample selection is to choose a batch of K sentences that maximizes the cumulative benefit value of n-grams covered by the chosen sentences. The contribution of each sentence towards the objective function is equal to the sum of benefit values of all unique n-grams it contains.

The sentence selection problem is closely related to the classical set covering problem, one of 21 NP-complete problems described in Karp's

seminal paper on reducibility (Karp, 1972). The decision version of set covering is as follows: given a set of elements T and a set of subsets within T, say $\mathcal{T} = \{T_j \subseteq T\}$, is it possible to select k subsets such that their union is the superset T? This problem is known to be NP-Complete.

To visualize the similarity of set covering to the sentence selection problem, note that the elements of T are akin to n-grams, and the subsets T_j's correspond to sentences. It is easy to show that the set covering problem can be reduced to the sentence selection problem. Thus, there does not exist a polynomial time algorithm for optimal sentence selection unless $P = NP$.

There is a standard greedy approximation algorithm for the minimum set covering problem (Cormen et al., 2001): at each iteration, choose the subset T_j that has the largest number of as yet uncovered elements of T. The procedure is repeated until all elements in T are covered. We next present a variant of this algorithm for the sentence selection problem. Our algorithm address two differences between sentence selection and set covering: (a) each n-gram provides a distinct benefit value on covering, and (b) we must select only K sentences and maximize the cumulative benefit.

4 Greedy Sample Selection

The greedy solution constructs batches iteratively by choosing, at each step, the sentence whose total current benefit is the largest. Each sentence in the candidate pool is decomposed into its constituent n-grams, and the sum of benefit values of these n-grams is computed. The sentence that scores highest on this criterion is chosen. Resetting the benefit values of n-grams in previously chosen sentences ensures diversity. The greedy selection technique is illustrated in Algorithm 1. Each iteration of the master loop selects one sentence based on the local maximum of the objective function. The indicator function $\mathcal{I}_i(\cdot)$ returns unity if n-gram n_i is present in the argument sentence, and zero otherwise.

The greedy algorithm provides a highly-scalable approximation to the solution and can be applied to systems with millions of n-grams and candidate sentences. It is, however, sub-optimal in general; potentially better solutions can be obtained by casting it in an integer linear programming (ILP) framework, as discussed below.

Algorithm 1 Greedy Sample Selection

$\mathbf{B} \leftarrow ()$
for $k = 1$ to K **do**
  $p^* \leftarrow \arg\max_{p \in \mathbf{P}} \sum_{i=1}^{m} b_i \mathcal{I}_i(p)$
  $B(k) \leftarrow p^*$
  $\mathbf{P} \leftarrow \mathbf{P} - \{p^*\}$
  $b_i \leftarrow 0 \ \forall \{i \mid \mathcal{I}_i(p^*) = 1\}$
end for
return $\mathbf{B}$

5 Integer Linear Programming (ILP)

We define a set of indicator variables, x_j, for the sentences, $x_j = 1$ if sentences p_j is selected and 0 otherwise. Similarly, there are a set of indicator variables, y_i, for the n-grams, $y_i = 1$ if n-gram n_i is covered by some selected sentence and 0 otherwise. Sentence selection can be expressed as the following integer linear program (ILP):

$$\begin{aligned} \max: \ & \sum_i b_i y_i \\ \text{subj. to.}: \ & y_i \leq \sum_{j \mid \mathcal{I}_i(p_j)=1} x_j \ \ \forall\, i \\ & \sum_j x_j \leq K \\ & 0 \leq y_i \leq 1 \ \ \forall\, i \ \ , \ \ x_j \in \{0,1\} \ \ \forall\, j \end{aligned} \tag{1}$$

Notice that since $b_i \geq 0$, in order to maximize the optimization function each y_i will be set to 1 whenever at least one of the sentence covering its n-gram is selected, $x_j = 1$.

In general, exact optimization of ILP is NP-Hard. However, there are several publicly available solvers for ILPs with thousands of variables. For instance, the open-source *lp-solve* program uses the Branch-and-Bound technique to solve ILPs and in our experiments handles systems with thousands of sentences.

5.1 Including Application Constraints

Unlike greedy selection, ILP allows us to impose additional application or domain specific global constraints within the optimization framework. One example is to bound the total number of words in the selected sentences rather than number of chosen sentences. This is useful because when the number of sentences is constrained, the system is biased to choose longer sentences as they would cover more n-grams. Assuming manual translation cost to be linear in the number of words, we

can put a bound on the total length of the chosen sentences. Let l_j be the length of sentence p_j. We can put a bound $\sum_j l_j x_j \leq L$. Imposing such a bound is similar to the Knapsack problem, a standard NP-Complete problem (Cormen et al., 2001). We can also incorporate prior information on the goodness of sentences by including sentence costs (e.g. syntactic well-formedness, length, etc.), c_j's, within the optimization function: $\sum_i b_i y_i + \sum_j c_j x_j$. In contrast to the greedy algorithm, ILP provides a natural framework to perform *joint* optimization over multiple types of constraints.

5.2 Solving the ILP for Practical Problems

Even moderate size SMT applications involve tens of thousands of n-grams and sentences, e.g., one of our test conditions has approximately $79,000$ n-grams and $100,000$ sentences. Each sentence would result in an integer variable in the ILP. State-of-the-art solvers have difficulty handling such large problems, e.g., *lp-solve* was unable to solve the ILP for a system with $100,000$ sentences. We propose a two-step solution to achieve scalability. If k sentences must be selected in a given active learning iteration, we use the greedy algorithm to prune the problem by choosing $k' > k$ sentences from the corpus, and subsequently construct the ILP on this smaller problem to select the required k sentences. While ILP is optimal, greedily pruning the problem for ILP may result in sub-optimality.

We observed the run time and optimization value computed by ILP for different k', keeping k constant at 16. We chose smaller (k', k) for these simulations to allow ILP to run to completion. Figure 2 summarizes our findings. Note that the optimum improves with larger prune size and requires more iterations. Larger prune sizes allow the ILP to choose from a larger pool of sentences, and are therefore likely to improve the optimum. However, these typically require more iterations. In this paper, we used the greedy algorithm to prune the problem to $K' = 1000$ sentences, and then select $K = 400$ sentences using ILP (we restricted ILP run-time to 20 minutes per batch).

6 Experimental Results

We demonstrate the effectiveness of greedy and ILP-based sample selection by conducting simulation experiments on two resource-poor language pairs commissioned under the DARPA Transtac speech-to-speech translation initiative, viz. English-Pashto (E2P) and English-Dari (E2D). In both cases, only a small fraction of the available training data is pertinent to the translation task. This simulates a condition where a large source language corpus (e.g. English) is harvested from the Web, of which only a small fraction is relevant to the target SMT system.

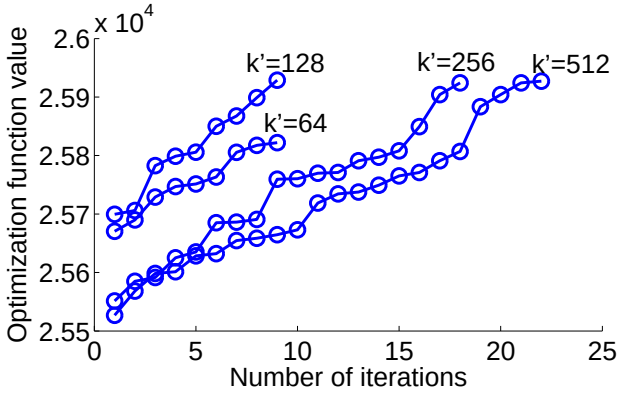


Figure 2: Optimization function value for different iterations of ILP branch-and-bound, for different sizes of pruned problems ($k' = 64, 128, 256, 512$) for constant choice cardinality, $k = 16$.

We simulate low-resource conditions by sequestering the majority of available parallel training data. A seed translation model is bootstrapped with a very small subset of the training corpus; source sentences of the remainder constitute the candidate pool. Because obtaining monolingual text in the target language is usually not a constraint, we train the target language model (LM) from all available target language sentences in the training corpus.

We then apply the proposed selection algorithms to choose fixed-size batches from the pool. Translations for the selected sentences are obtained from the sequestered parallel corpus (thus simulating a human oracle). The chosen batch and its translation is appended to the seed corpus $\mathbf{S}$ for retraining the SMT. At each iteration, we independently decode the test set and evaluate translation accuracy in order to compare the trajectory of BLEU for these and other competing selection strategies:

- *Random*: Source sentences are uniformly sampled from the candidate pool $\mathbf{P}$.
- *Dissimilarity*: Select sentences from $\mathbf{P}$ with the largest number of n-grams not seen in $\mathbf{S}$

(Eck et al., 2005; Haffari et al., 2009).

- *Longest*: Pick the longest sentences from the candidate pool $\mathbf{P}$.
- *Discriminative*: Choose sentences that potentially minimize translation error using a maximum-entropy pairwise comparator (Ananthakrishnan et al., 2010).
- *Greedy*: Simple greedy selection with proposed error-projection benefit objective.
- *ILP*: Integer linear programming optimization with error-projection benefit objective.

English-to-Pashto Simulation: The 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. In addition, DARPA has made available to all Transtac participants an open 564-sentence E2P test set with four target references for each input.

We trained a baseline E2P SMT system from all available E2P and reversed P2E data. The full-system BLEU scores on the single-reference internal test set and on the multi-reference DARPA evaluation test set were 10.8 and 24.4, respectively. We set up active learning simulation by randomly sampling 1,000 sentence pairs from the directional E2P training partition to obtain the seed training corpus. The remainder of this set, and the entire reversed P2E training partition were combined to create the pool. The reversed directional P2E data is considered irrelevant as far as the E2P test sets are concerned. The pool thus consists of 30% in-domain and 70% irrelevant sentence pairs. We simulated 35 iterations with batches of 400 sentences each; the seed corpus grows to 15,000 sentence pairs at the end of the simulation.

English-to-Dari Simulation: The E2D data is also derived from a two-way collection of spoken dialogues. The directional E2D training, development, and test sets consist of 11.6k, 3.2k, and 2.8k sentence pairs, respectively. The directional D2E training set consists of 52.9k sentence pairs. The full-system BLEU on the E2D test set was 15.1.

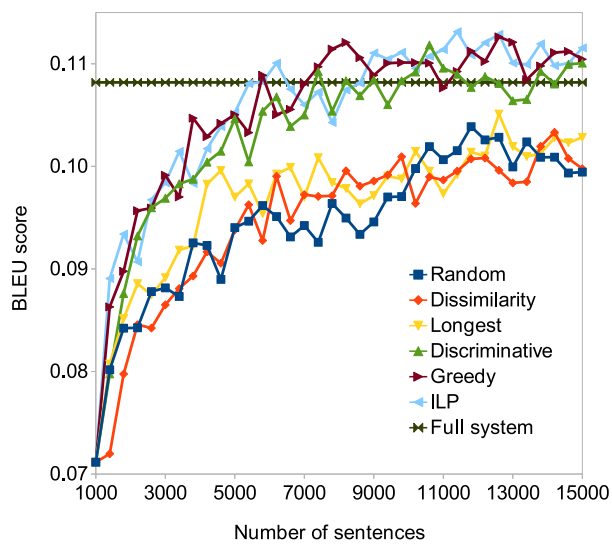
As with E2P, the seed training corpus was obtained by randomly sampling 1,000 sentence pairs from the directional E2D training partition. All remaining parallel training data were designated as the candidate pool. Thus, only about 17% of the candidate pool is considered relevant with respect to the E2D test set. Again, we simulated 35 iterations with batches of 400 sentences each.

BLEU Trajectories: The trajectories of BLEU scores for the E2P and E2D test sets are shown in Figures 3(a), 3(b), and 3(c), respectively. The horizontal line near the top of each plot represents the corresponding full-system BLEU score. The following observations are noteworthy:

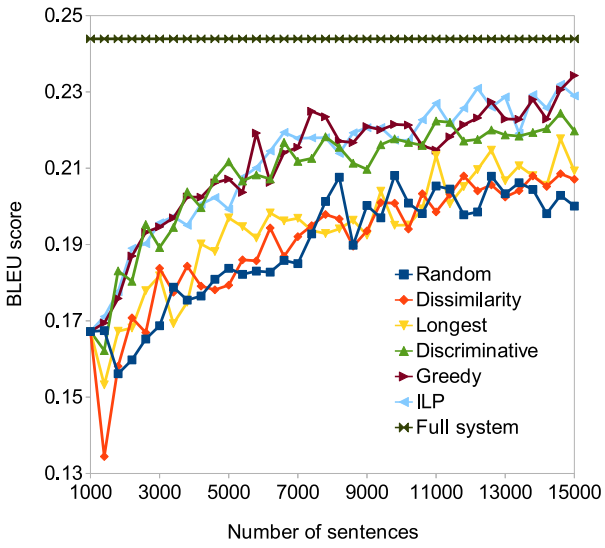
- BLEU scores using the proposed greedy and ILP-based selection methods ramp up very quickly to the full-system level using a small fraction of available training data. On the E2P single-reference test set, the top-line BLEU score of 10.8 is attained after just 12 iterations (5.8k sentence pairs), as against 100k sentence pairs for the full system (**6%** of the corpus). Likewise for E2D, the full-system BLEU score of 15.1 is attained after only 18 iterations (8.2k sentence pairs), as opposed to 65k sentence pairs for the full system (**13%** of the corpus).
- In some cases, BLEU scores with training corpora constructed using active learning exceed those obtained with the full system. This is because our selection algorithms are biased to choose relevant, in-domain sentences from the candidate pool. Initially, the training corpus is kept free of outliers that cause performance degradation in the full system. With more iterations of selection, the latter eventually find their way into the training set, causing translation performance to settle around the top-line BLEU scores.
- Under identical initial conditions at the first iteration of active learning, the ILP benefit optimum exceeds the greedy optimum by 444.5 units for E2P, and by 420.3 units for E2D. This confirms the theoretical superiority of ILP over greedy selection.

We computed total area under the BLEU curves for the various selection techniques as a single figure of merit. Summarized in Table 1, the BLEU-Iteration product shows source error-projection with the ILP selection algorithm outperforming all competing techniques, including discriminative sample selection (Ananthakrishnan et al., 2010).

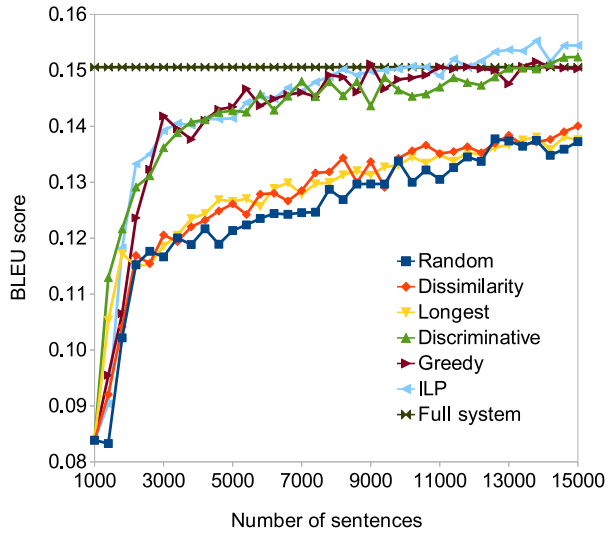
Global Length Constraint: The above simula-



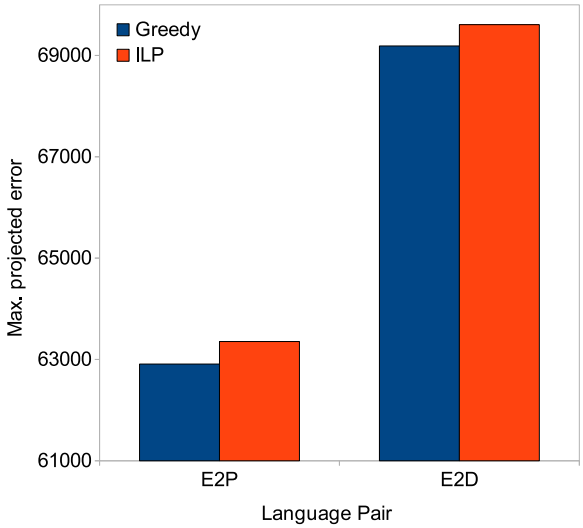
(a) Trajectory of BLEU (E2P-SingleRef)



(b) Trajectory of BLEU (E2P-MultiRef)



(c) Trajectory of BLEU (E2D-SingleRef)



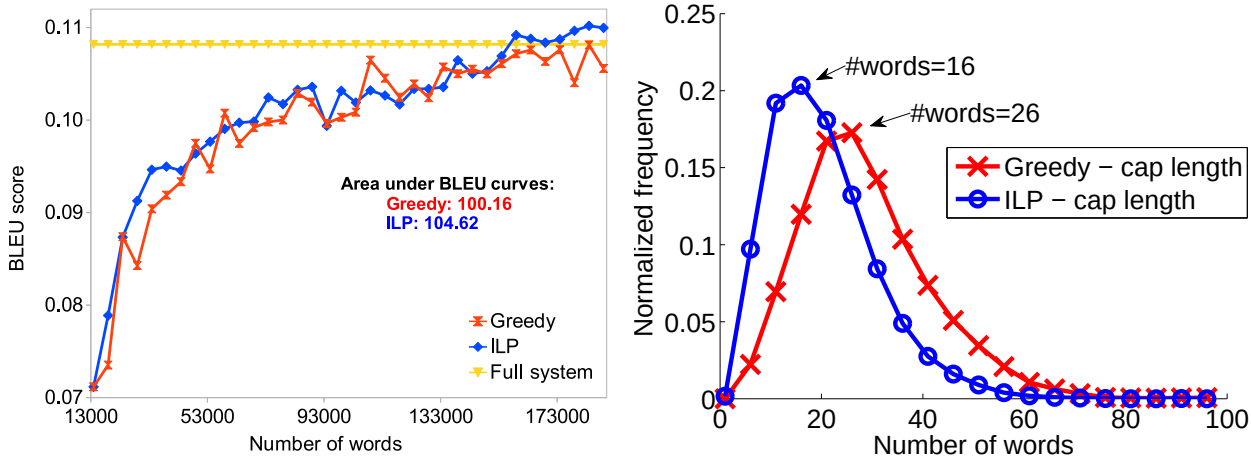
(d) Maximum benefit for greedy and ILP selection

Figure 3: Simulation results for E2P and E2D.

Method	E2P-SR	E2P-MR	E2D-SR
Random	83.03	78.76	144.65
Dissim.	82.34	81.18	154.17
Longest	89.89	93.65	155.72
Discrim.	113.54	145.32	205.56
Greedy	120.83	155.64	204.95
ILP	**121.16**	**156.70**	**210.51**

Table 1: Area under the BLEU curve with respect to 0-th iteration baseline.

tions choose a fixed number of sentences at each iteration. However, the ILP optimization framework also permits the integration of global constraints, such as the number of words in a batch. This is an important practical benefit, as most professional translators charge by word rather than by sentence. Other selection methods can only support such constraints in an ad-hoc fashion. To evaluate this feature of ILP selection, we implemented a variant of the greedy algorithm where the stopping criterion is number of words selected. We then imposed a global length constraint for ILP as described in Section 5.1, and compared BLEU scores across 35 simulation iterations for E2P with a limit of 5,000 words per batch (Figure 4(a)). The trend in BLEU indicates that ILP provides a better framework for integration of such constraints.



(a) Length-capped BLEU trajectory (E2P-SingleRef). Full system training data contains 1.46M source words.

(b) Normalized histogram of sentence length

Figure 4: Comparison of greedy and ILP with global length constraint of 5,000 words per batch.

Moreover, the overall sentence length distribution shown in Figure 4(b) indicates that the greedy algorithm prefers a smaller number of very long sentences, whereas ILP prefers a larger number of shorter sentences. The latter is an important practical benefit for applications such as crowdsourcing, because annotators often find it difficult to translate long sentences. Secondly, automatic word alignment quality for long sentence pairs is often poor. Finally, larger number of sentences in the training pool is likely to increase diversity.

7 Discussion and Future Work

Active learning provides a useful framework for alleviating the significant costs associated with developing SMT systems for resource-poor language pairs. In this paper, we introduced a novel criterion for active sample selection, viz. back-projected translation error. Candidate instances that score well on this criterion are chosen for translation by a bilingual human expert. We showed that the problem of maximizing the error-projection objective function is closely related to the set-covering problem, known to be NP-complete.

We used a simple greedy selection algorithm as a first approximation to the solution. BLEU trajectories from simulation experiments demonstrated the superiority of this scheme to competing active learning algorithms. We then proposed an optimization framework to maximize the objective function for sample selection, and provided a solution via ILP. The ILP-based selection algorithm also supports global constraints on the optimization problem, e.g. overall corpus size in words, which neither greedy selection nor other competing strategies can implement in a principled fashion. We also showed that ILP was superior to the greedy approach when constraining selected batches by total number of words, rather than by number of sentences.

The proposed approach is shown to outperform competing active learning strategies when the candidate pool contains a small number of high-impact samples buried within a large corpus of mostly irrelevant text. Guided by an in-domain development set, our approach always selects relevant samples that are likely to provide maximal benefit to the SMT system. Pilot experiments suggest that this approach may not be as effective when the candidate pool is completely task-relevant. Another weakness of our technique is the reliance on a relatively small development set to guide selection. Performance may saturate once all sentences in the development set can be translated accurately. Generally, a large, rich development set will tend to give better results. In the future, we plan to experiment with random sampling to vary the development set at each active learning iteration.

We have shown the effectiveness of our approach in the context of phrase-based SMT systems. The same principles can be extended to hierarchical or syntax-based SMT architectures with minimal effort.

References

Sankaranarayanan Ananthakrishnan, Rohit Prasad, David Stallard, and Prem Natarajan. 2010. Discriminative sample selection for statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, EMNLP '10, pages 626–635, Morristown, NJ, USA. Association for Computational Linguistics.

Michael Bloodgood and Chris Callison-Burch. 2010. Bucking the trend: large-scale cost-focused active learning for statistical machine translation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 854–864, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Thomas H. Cormen, Charles E. Lerserson, and Ronald L. Rivest. 2001. *Introduction to Algorithms*. Prentice Hall of India.

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 and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

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.

Richard M. Karp. 1972. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum.

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.

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.

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*, pages 160–167, Morristown, NJ, USA. Association for Computational Linguistics.

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*, page 589, 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.