A Semi-Supervised Batch-Mode Active Learning Strategy for Improved Statistical Machine Translation

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

Abstract

The availability of substantial, in-domain parallel corpora is critical for the development of high-performance statistical machine translation (SMT) systems. Such corpora, however, are expensive to produce due to the labor intensive nature of manual translation. We propose to alleviate this problem with a novel, semi-supervised, batch-mode *active learning* strategy that attempts to maximize in-domain coverage by selecting sentences, which represent a balance between domain match, translation difficulty, and batch diversity. Simulation experiments on an English-to-Pashto translation task show that the proposed strategy not only outperforms the random selection baseline, but also traditional active learning techniques based on dissimilarity to existing training data. Our approach achieves a relative improvement of 45.9% in BLEU over the seed baseline, while the closest competitor gained only 24.8% with the same number of selected sentences.

1 Introduction

Rapid development of statistical machine translation (SMT) systems for resource-poor language pairs is a problem of significant interest to the research community in academia, industry, and government. Tight turn-around schedules, budget restrictions, and scarcity of human translators preclude the production of large parallel corpora, which form the backbone of SMT systems.

Given these constraints, the focus is on making the best possible use of available resources. This usually involves some form of prioritized data collection. In other words, one would like to construct the smallest possible parallel training corpus that achieves a desired level of performance on unseen test data.

Within an *active learning* framework, this can be cast as a data selection problem. The goal is to choose, for manual translation, the most informative instances from a large *pool* of source language sentences. The resulting sentence pairs, in combination with any existing in-domain *seed* parallel corpus, are expected to provide a significantly higher performance gain than a naïve random selection strategy. This process is repeated until a certain level of performance is attained.

Previous work on active learning for SMT has focused on unsupervised dissimilarity measures for sentence selection. Eck et al. (2005) describe a selection strategy that attempts to maximize coverage by choosing sentences with the highest proportion of previously unseen n-grams. However, if the pool is not completely in-domain, this strategy may select irrelevant sentences, whose translations are unlikely to improve performance on an in-domain test set. They also propose a technique, based on TF-IDF, to de-emphasize sentences similar to those that have already been selected. However, this strategy is bootstrapped by random initial choices that do not necessarily favor sentences that are difficult to translate. Finally, they work exclusively with the source language and do not use any SMT-derived features to guide selection.

Haffari et al. (2009) propose a number of features, such as similarity to the seed corpus, translation probability, relative frequencies of n-grams and "phrases" in the seed vs. pool data, etc., for active learning. While many of their experiments use the above features independently to compare their relative efficacy, one of their experiments attempts to predict a rank, as a linear combination of these features, for each candidate sentence. The top-ranked sentences are chosen for manual translation. The latter strategy is particularly relevant to this paper, because the goal of our active

Proceedings of the Fourteenth Conference on Computational Natural Language Learning, pages 126–134,
Uppsala, Sweden, 15-16 July 2010.

learning strategy is not to compare features, but to learn the trade-off between various characteristics of the candidate sentences that potentially maximizes translation improvement.

The parameters of the linear ranking model proposed by Haffari et al. (2009) are trained using two held-out development sets $\mathbf{D_1}$ and $\mathbf{D_2}$ - the model attempts to learn the ordering of $\mathbf{D_1}$ that incrementally maximizes translation performance on $\mathbf{D_2}$. Besides the need for multiple parallel corpora and the computationally intensive nature of incrementally retraining an SMT system, their approach suffers from another major deficiency. It requires that the pool have the same distributional characteristics as the development sets used to train the ranking model. Additionally, they select all sentences that constitute a batch in a single operation following the ranking procedure. 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. This results in under-utilization of human translation resources.

In this paper, we propose a novel batch-mode active learning strategy that ameliorates the above issues. Our semi-supervised learning approach combines a parallel ranking strategy with several features, including domain representativeness, translation confidence, and batch diversity. The proposed approach includes a greedy, incremental batch selection strategy, which encourages diversity and reduces redundancy. The following sections detail our active learning approach, including the experimental setup and simulation results that clearly demonstrate its effectiveness.

2 Active Learning Paradigm

Active learning has been studied extensively in the context of multi-class labeling problems, and theoretically optimal selection strategies have been identified for simple classification tasks with metric features (Freund et al., 1997). However, natural language applications such as SMT present a significantly higher level of complexity. For instance, SMT model parameters (translation rules, language model n-grams, etc.) are not fixed in number or type, and vary depending on the training instances. This gives rise to the concept of *domain*. Even large quantities of out-of-domain training data usually do not improve translation performance. As we will see, this causes simple active selection techniques based on dissimilarity or translation difficulty to be ineffective, because they tend to favor out-of-domain sentences.

Our proposed active learning strategy is motivated by the idea that the chosen sentences should maximize coverage, and by extension, translation performance on an unseen test set. It should pick sentences that represent the target domain, while simultaneously enriching the training data with hitherto unseen, difficult-to-translate constructs that are likely to improve performance on a test set. We refer to the former as *representativeness* and to the latter as *difficulty*.

Since it is computationally prohibitive to retrain an SMT system for individual translation pairs, a batch of sentences is usually selected at each iteration. We desire that each batch be sufficiently *diverse*; this increases the number of concepts (phrase pairs, translation rules, etc.) that can be learned from manual translations of a selected batch. Thus, our active learning strategy attempts, at each iteration, to select a batch of mutually diverse source sentences, which, while introducing new concepts, shares at least some commonality with the target domain. This is done in a completely statistical, data-driven fashion.

In designing this active learning paradigm, we make the following assumptions.

- A small seed parallel corpus $\mathbf{S}$ is available for training an initial SMT system. This may range from a few hundred to a few thousand sentence pairs.
- Sentences must be selected from a large pool $\mathbf{P}$. This may be an arbitrary collection of in- and out-of-domain source language sentences. Some measure of redundancy is permitted and expected, i.e. some sentences may be identical or very similar to others.
- A development set $\mathbf{D}$ is available to tune the SMT system and train the selection algorithm. An unseen test set $\mathbf{T}$ is used to evaluate it.
- The seed, development, and test sets are derived from the target domain distribution.

To re-iterate, we do not assume or require the pool to have the same domain distribution as the seed, development, and test sets. This reflects a real-world scenario, where the pool may be drawn from multiple sources (e.g. targeted collections, newswire text, web, etc.). This is a key departure from existing work on active learning for SMT.

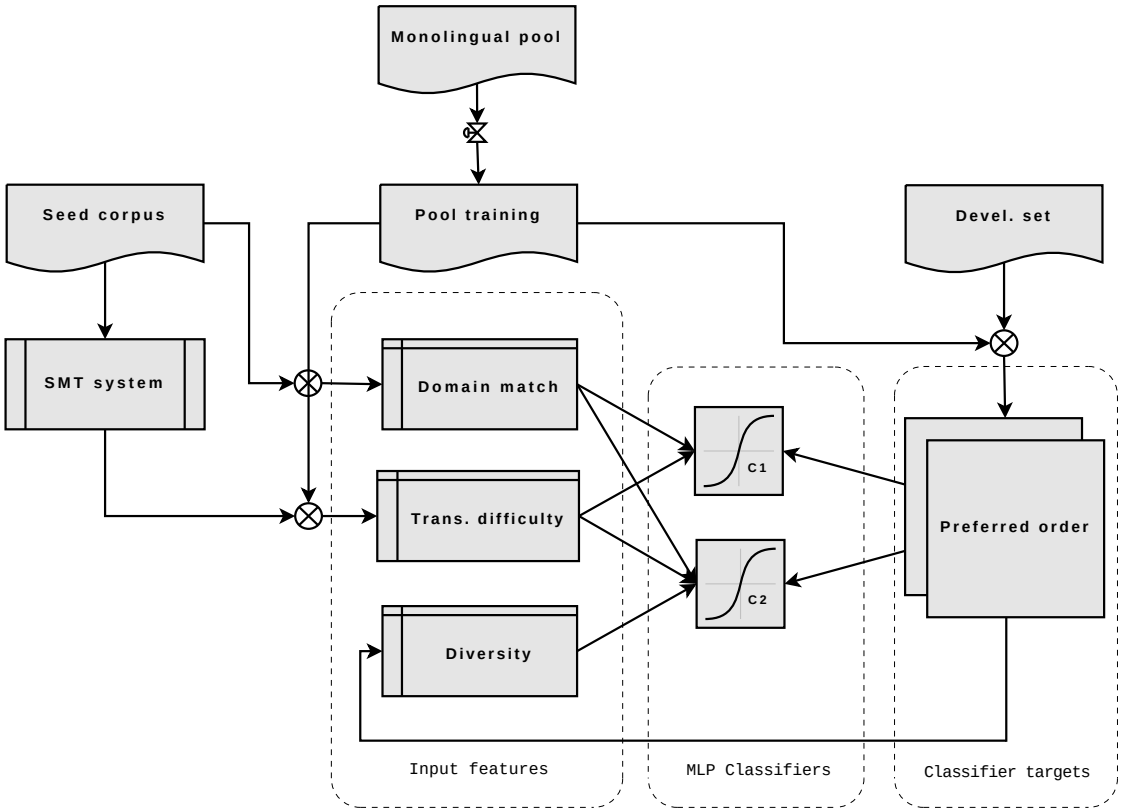


Figure 1: Flow-diagram of the active learner.

3 Active Learning Architecture

Figure 1 illustrates the proposed active learning architecture in the form of a high-level flow-diagram. We begin by randomly sampling a small fraction of the large monolingual pool $\mathbf{P}$ to create a *pool training set* $\mathbf{P_T}$, which is used to train the learner. The remainder, which we call the *pool evaluation set* $\mathbf{P_E}$, is set aside for active selection. We also train an initial phrase-based SMT system (Koehn et al., 2003) with the available seed corpus. The pool training set $\mathbf{P_T}$, in conjunction with the seed corpus $\mathbf{S}$, initial SMT system, and held-out development set $\mathbf{D}$, is used to derive a number of input features as well as target labels for training two parallel classifiers.

3.1 Preferred Ordering

The learner must be able to map input features to an ordering of the pool sentences that attempts to maximize coverage on an unseen test set. We teach it to do this by providing it with an ordering of $\mathbf{P_T}$ that incrementally maximizes source coverage on $\mathbf{D}$. This *preferred ordering* algorithm incrementally maps sentences in $\mathbf{P_T}$ to a ordered set $\mathbf{O_T}$ by picking, at each iteration, the sentence with the highest *coverage criterion* with respect to $\mathbf{D}$, and inserting it at the current position within $\mathbf{O_T}$. The coverage criterion is based on content-word n-gram overlap with $\mathbf{D}$, discounted by constructs already observed in $\mathbf{S}$ and higher-ranked sentences in $\mathbf{O_T}$, as illustrated in Algorithm 1. Our hypothesis is that sentences, which maximally improve coverage, likely lead to bigger gains in translation performance as well.

The $O(|\mathbf{P_T}|^2)$ complexity of this algorithm is one reason we restrict $\mathbf{P_T}$ to a few thousand sentences. Another reason not to order the entire pool and simply select the top-ranked sentences, is that batches thus constructed would overfit the development set on which the ordering is based, and not generalize well to an unseen test set.

3.2 Ranker Features

Each candidate sentence in the pool is represented by a vector of features, which fall under one of the three categories, viz. representativeness, difficulty, and diversity. We refer to the first two as *context-independent*, because they can be computed independently for each sentence. Diversity is a *context-dependent* feature and must be evaluated in the context of an ordering of sentences.

Algorithm 1 Preferred ordering

$\mathbf{O_T} \leftarrow ()$
$S_g \leftarrow \text{count}(g) \quad \forall g \in \text{ngr}(\mathbf{S})$
$D_g \leftarrow \text{count}(g) \quad \forall g \in \text{ngr}(\mathbf{D})$
for $k = 1$ to $|\mathbf{P_T}|$ **do**
  $\mathbf{P_U} \leftarrow \mathbf{P_T} - \mathbf{O_T}$
  $y^* \leftarrow \arg\max_{y \in \mathbf{P_U}} \sum_{g \in \text{ngr}(y)} \frac{y_g \times D_g \times n}{S_g + 1}$
  $O_T(k) \leftarrow y^*$
  $S_g \leftarrow S_g + y^*_g \quad \forall g \in \text{ngr}(y^*)$
end for
return $\mathbf{O_T}$

3.2.1 Domain Representativeness

Domain representativeness features gauge the degree of similarity between a candidate pool sentence and the seed training data. We quantify this using an n-gram overlap measure between candidate sentence x and the seed corpus $\mathbf{S}$ defined by Equation 1.

$$sim(x, \mathbf{S}) = \frac{\sum_{g \in \text{ngr}(x)} x_g \times \frac{min(S^n_g, C_n)}{C_n}}{\sum_{g \in \text{ngr}(x)} x_g} \quad (1)$$

x_g is the number of times n-gram g occurs in x, S_g the number of times it occurs in the seed corpus, n its length in words, and C_n the count of n-grams of length n in $\mathbf{S}$. Longer n-grams that occur frequently in the seed receive high similarity scores, and vice-versa. In evaluating this feature, we only consider n-grams up to length five that contain least one content word.

Another simple domain similarity feature we use is sentence length. Sentences in conversational domains are typically short, while those in web and newswire domains run longer.

3.2.2 Translation Difficulty

All else being equal, the selection strategy should favor sentences that the existing SMT system finds difficult to translate. To this end, we estimate a confidence score for each SMT hypothesis, using a discriminative classification framework reminiscent of Blatz et al. (2004). Confidence estimation is treated as a binary classification problem, where each hypothesized word is labeled "*correct*" or "*incorrect*". Word-level reference labels for training the classifier are obtained from Translation Edit Rate (TER) analysis, which produces the lowest-cost alignment between the hypotheses and the gold-standard references (Snover et al., 2006). A hypothesized word is "*correct*" if it aligns to itself in this alignment, and "*incorrect*" otherwise.

We derive features for confidence estimation from the phrase derivations used by the decoder in generating the hypotheses. For each target word, we look up the corresponding source phrase that produced it, and use this information to compute a number of features from the translation phrase table and target language model (LM). These include the in-context LM probability of the target word, the forward and reverse phrase translation probabilities, the maximum forward and reverse word-level lexical translation probabilities, number of competing target phrases in which the target word occurs, etc. In all, we use 11 word-level features (independent of the active learning features) to train the classifier in conjunction with the abovementioned binary reference labels.

A logistic regression model is used to directly estimate the posterior probability of the binary word label. Thus, our confidence score is essentially the probability of the word being "*incorrect*". Sentence-level confidence is computed as the geometric average of word-level posteriors. Confidence estimation models are trained on the held-out development set.

We employ two additional measures of translation difficulty for active learning: (a) the number of "unknown" words in target hypotheses caused by untranslatable source words, and (b) the average length of source phrases in the 1-best SMT decoder derivations.

3.2.3 Batch Diversity

Batch diversity is evaluated in the context of an explicit ordering of the candidate sentences. In general, sentences that are substantially similar to those above them in a ranked list have low diversity, and vice-versa. We use content-word n-gram overlap to measure similarity with previous sentences, per Equation 2.

$$d(b \mid \mathbf{B}) = 1.0 - \frac{\sum_{g \in \text{ngr}(b)} n \times B_g}{\sum_{g \in \text{ngr}(b)} n \times max(B_g, 1.0)} \quad (2)$$

$\mathbf{B}$ represents the set of sentences ranked higher than the candidate b, for which we wish to evaluate diversity. B_g is the number of times n-gram g

occurs in $\mathbf{B}$. Longer, previously unseen n-grams serve to boost diversity. The first sentence in a given ordering is always assigned unit diversity. The coverage criterion used by the preferred ordering algorithm in Section 3.1 ensures good correspondence between the rank of a sentence and its diversity, i.e. higher-ranked in-domain sentences have higher diversity, and vice-versa.

3.3 Training the Learner

The active learner is trained on the pool training set $\mathbf{P_T}$. The seed training corpus $\mathbf{S}$ serves as the basis for extracting domain similarity features for each sentence in this set. Translation difficulty features are evaluated by decoding sentences in $\mathbf{P_T}$ with the seed SMT system. Finally, we compute diversity for each sentence in $\mathbf{P_T}$ based on its preferred order $\mathbf{O_T}$ according to Equation 2. Learning is *semi-supervised* as it does not require translation references for either $\mathbf{P_T}$ or $\mathbf{D}$.

Traditional ranking algorithms such as PRank (Crammer and Singer, 2001) work best when the number of ranks is much smaller than the sample size; more than one sample can be assigned the same rank. In the active learning problem, however, each sample is associated with a unique rank. Moreover, the dynamic range of ranks in $\mathbf{O_T}$ is significantly smaller than that in $\mathbf{P_E}$, to which the ranking model is applied, resulting in a mismatch between training and evaluation conditions.

We overcome these issues by re-casting the ranking problem as a binary classification task. The top 10% sentences in $\mathbf{O_T}$ are assigned a "*select*" label, while the remaining are assigned a contrary "*do-not-select*" label. The input features are mapped to class posterior probabilities using multi-layer perceptron (MLP) classifiers. The use of posteriors allows us to assign a unique rank to each candidate sentence. The best candidate sentence is the one to which the classifier assigns the highest posterior probability for the "*select*" label. We use one hidden layer with eight sigmoid-activated nodes in this implementation.

Note that we actually train two MLP classifiers with different sets of input features as shown in Figure 1. Classifier $\mathcal{C}_1$ is trained using only the context-independent features, whereas $\mathcal{C}_2$ is trained with the full set of features including batch diversity. These classifiers are used to implement an incremental, greedy selection algorithm with parallel ranking, as explained below.

Algorithm 2 Incremental greedy selection

$\mathbf{B} \leftarrow ()$
for $k = 1$ to N **do**
  $\mathbf{P_{ci}} \leftarrow \{x \in \mathbf{P_E} \mid d(x \mid \mathbf{B}) = 1.0\}$
  $\mathbf{P_{cd}} \leftarrow \{x \in \mathbf{P_E} \mid d(x \mid \mathbf{B}) < 1.0\}$
  $\mathbf{C} \leftarrow \mathcal{C}_1(f_{ci}(\mathbf{P_{ci}})) \cup \mathcal{C}_2(f_{cd}(\mathbf{P_{cd}}, \mathbf{B}))$
  $b_k \leftarrow \arg\max_{x \in \mathbf{P_E}} \mathbf{C}(x)$
  $\mathbf{P_E} \leftarrow \mathbf{P_E} - \{b_k\}$
end for
return $\mathbf{B}$

4 Incremental Greedy Selection

Traditional rank-and-select batch construction approaches choose constituent sentences independently, and therefore cannot ensure that the chosen sentences are sufficiently diverse. Our strategy implements a greedy selection algorithm that constructs each batch iteratively; the decision b_k (the sentence to fill the k^{th} position in a batch) depends on all previous decisions $b_1, \cdots, b_{k-1}$. This allows de-emphasizing sentences similar to those that have already been placed in the batch, while favoring samples containing previously unseen constructs.

4.1 Parallel Ranking

We begin with an empty batch $\mathbf{B}$, to which sentences from the pool evaluation set $\mathbf{P_E}$ must be added. We then partition the sentences in $\mathbf{P_E}$ in two mutually-exclusive groups $\mathbf{P_{cd}}$ and $\mathbf{P_{ci}}$. The former contains candidates that share at least one content-word n-gram with any existing sentences in $\mathbf{B}$, while the latter consists of sentences that do not share any overlap with them. Note that $\mathbf{B}$ is empty to start with; thus, $\mathbf{P_{cd}}$ is empty and $\mathbf{P_{ci}} = \mathbf{P_E}$ at the beginning of the first iteration of selection. The diversity feature is computed for each sentence in $\mathbf{P_{cd}}$ based on existing selections in $\mathbf{B}$, while the context-independent features are evaluated for sentences in both partitions.

Next, we apply $\mathcal{C}_1$ to $\mathbf{P_{ci}}$ and $\mathcal{C}_2$ to $\mathbf{P_{cd}}$ and independently obtain posterior probabilities for the "*select*" label for both partitions. We take the union of class posteriors from both partitions and select the sentence with the highest probability of the "*select*" label to fill the next slot b_k, corresponding to iteration k, in the batch. The selected sentence is subsequently removed from $\mathbf{P_E}$.

The above *parallel ranking* technique (Algorithm 2) is applied iteratively until the batch

reaches a pre-determined size N. At iteration k, the remaining sentences in $\mathbf{P_E}$ are partitioned based on overlap with previous selections $b_1, \cdots, b_{k-1}$ and ranked based on the union of posterior probabilities generated by the corresponding classifiers. This ensures that sentences substantially similar to those that have already been selected receive a low diversity score, and are suitably de-emphasized. Depending on the characteristics of the pool, batches constructed by this algorithm are likely more diverse than a simple rank-and-select approach.

5 Experimental Setup and Results

We demonstrate the effectiveness of the proposed sentence selection algorithm by performing a set of simulation experiments in the context of an English-to-Pashto (E2P) translation task. We simulate a low-resource condition by using a very small number of training sentence pairs, sampled from the collection, to bootstrap a phrase-based SMT system. The remainder of this parallel corpus is set aside as the pool.

At each iteration, the selection algorithm picks a fixed-size batch of source sentences from the pool. The seed training data are augmented with the chosen source sentences and their translations. A new set of translation models is then estimated and used to decode the test set. We track SMT performance across several iterations and compare the proposed algorithm to a random selection baseline as well as other common selection strategies.

5.1 Data Configuration

Our English-Pashto data originates from a two-way collection of spoken dialogues, and thus consists of two parallel sub-corpora: a directional E2P corpus and a directional Pashto-to-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.

We obtain a seed training corpus for the simulation experiments by randomly sampling 1,000 sentence pairs from the directional E2P training partition. The remainder of this set, and the entire reversed directional P2E training partition are 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, leaving us with 99.4k candidate sentences.

5.2 Selection Strategies

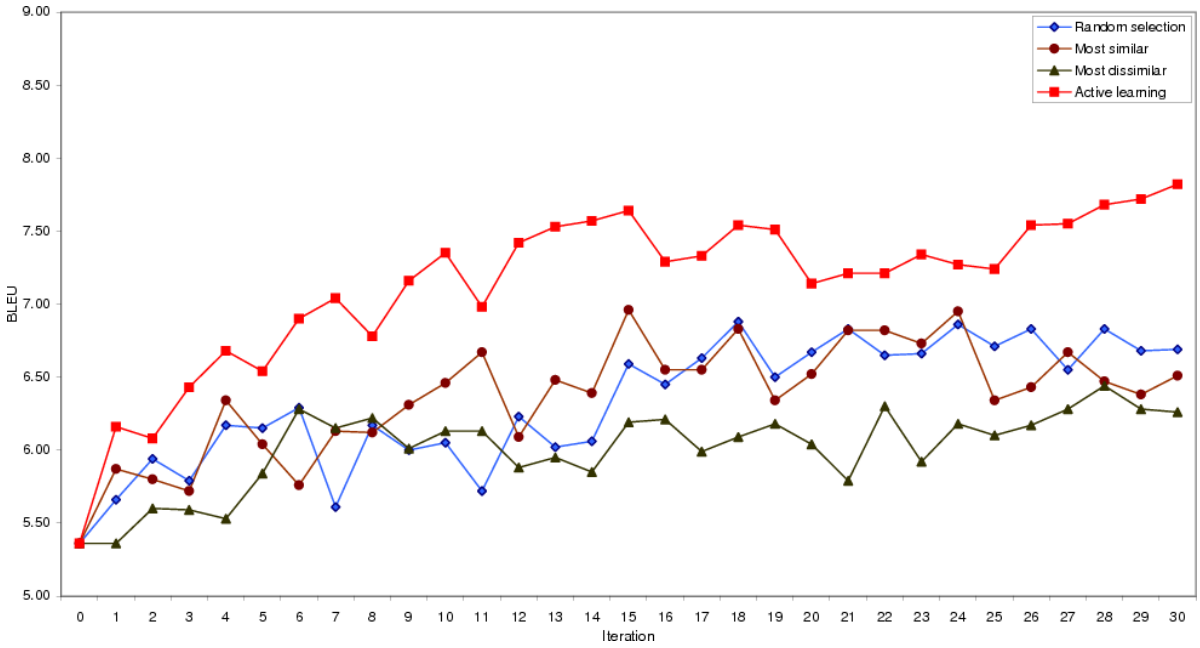
We implement the following strategies for sentence selection. In all cases, we use a fixed-size batch of 200 sentences per iteration.

- *Random selection*, in which source sentences are uniformly sampled from $\mathbf{P_E}$.
- *Similarity selection*, where we choose sentences that exhibit the highest content-word n-gram overlap with $\mathbf{S}$.
- *Dissimilarity selection*, which selects sentences having the lowest degree of content-word n-gram overlap with $\mathbf{S}$.
- *Active learning* with greedy incremental selection, using a learner to maximize coverage by combining various input features.

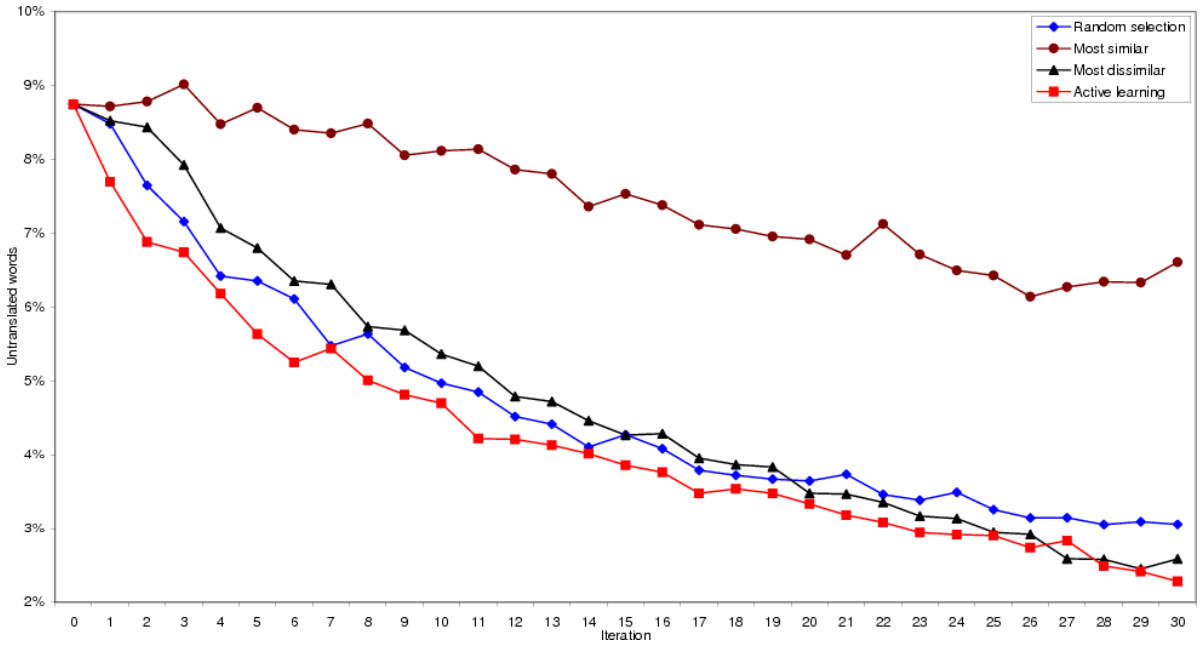
We simulate a total of 30 iterations, with the original 1,000 sample seed corpus growing to 7,000 sentence pairs.

5.3 Simulation Results

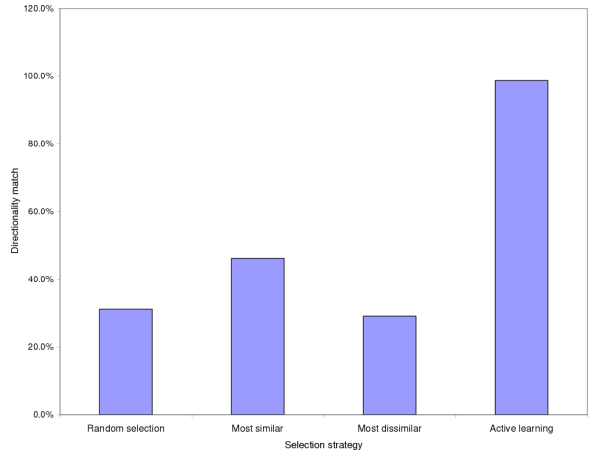
We track SMT performance at each iteration in two ways. The first and most effective method is to simply use an objective measure of translation quality, such as BLEU (Papineni et al., 2001). Figure 2(a) illustrates the variation in BLEU scores across iterations for each selection strategy. We note that the proposed active learning strategy performs significantly better at every iteration than random, similarity, and dissimilarity-based selection. At the end of 30 iterations, the BLEU score gained 2.46 points, a relative improvement of 45.9%. By contrast, the nearest competitor was the random selection baseline, whose performance gained only 1.33 points in BLEU, a 24.8% improvement. Note that we tune the phrase-based SMT feature weights using MERT (Och, 2003) once in the beginning, and use the same weights across all iterations. This allowed us to compare selection methods without variations introduced by fluctuation of the weights.



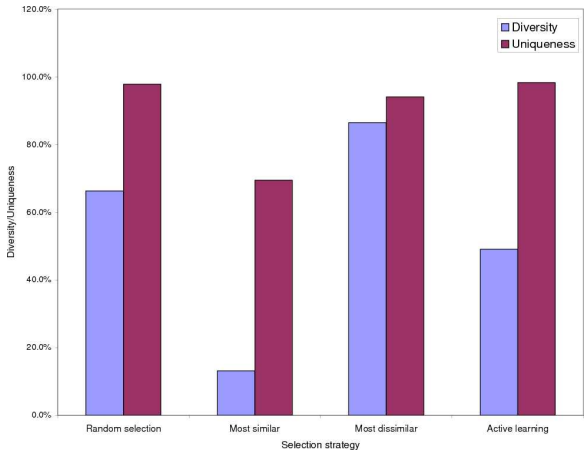
(a) Trajectory of BLEU



(b) Trajectory of untranslated word ratio



(c) Directionality match



(d) Diversity/Uniqueness

Figure 2: Simulation results for data selection. Batch size at each iteration is 200 sentences.

The second method measures test set coverage in terms of the proportion of untranslated words in the SMT hypotheses, which arise due to the absence of appropriate in-context phrase pairs in the training data. Figure 2(b) shows the variation in this measure for the four selection techniques. Again, the proposed active learning algorithm outperforms its competitors across nearly all iterations, with very large improvements in the initial stages. Overall, the proportion of untranslated words dropped from 8.74% to 2.28% after 30 iterations, while the closest competitor (dissimilarity selection) dropped to 2.59%.

It is also instructive to compare the distribution of the 6,000 sentences selected by each strategy at the end of the simulation to determine whether they came from the "in-domain" E2P set or the "out-of-domain" P2E collection. Figure 2(c) demonstrates that only 1.3% of sentences were selected from the reversed P2E set by the proposed active learning strategy. On the other hand, 70.9% of the sentences selected by the dissimilarity-based technique came from the P2E collection, explaining its low BLEU scores on the E2P test set. Surprisingly, similarity selection also chose a large fraction of sentences from the P2E collection; this was traced to a uniform distribution of very common sentences (e.g. "*thank you*", "*okay*", etc.) across the E2P and P2E sets.

Figure 2(d) compares the uniqueness and overall n-gram diversity of the 6,000 sentences chosen by each strategy. The similarity selector received the lowest score on this scale, explaining the lack of improvement in coverage as measured by the proportion of untranslated words in the SMT hypotheses. Again, the proposed approach exhibits the highest degree of uniqueness, underscoring its value in lowering batch redundancy.

It is interesting to note that dissimilarity selection is closest to the proposed active learning strategy in terms of coverage, and yet exhibits the worst BLEU scores. This confirms that, while there is overlap in their vocabularies, the E2P and P2E sets differ significantly in terms of longer-span constructs that influence SMT performance.

These results clearly demonstrate the power of the proposed strategy in choosing diverse, in-domain sentences that not only provide superior performance in terms of BLEU, but also improve coverage, leading to fewer untranslated concepts in the SMT hypotheses.

6 Conclusion and Future Directions

Rapid development of SMT systems for resource-poor language pairs requires judicious use of human translation capital. We described a novel active learning strategy that automatically learns to pick, from a large monolingual pool, sentences that maximize in-domain coverage. In conjunction with their translations, they are expected to improve SMT performance at a significantly faster rate than existing selection techniques.

We introduced two key ideas that distinguish our approach from previous work. First, we utilize a sample of the candidate pool, rather than an additional in-domain development set, to learn the mapping between the features and the sentences that maximize coverage. This removes the restriction that the pool be derived from the target domain distribution; it can be an arbitrary collection of in- and out-of-domain sentences.

Second, we construct batches using an incremental, greedy selection strategy with parallel ranking, instead of a traditional batch rank-and-select approach. This reduces redundancy, allowing more concepts to be covered in a given batch, and making better use of available resources.

We showed through simulation experiments that the proposed strategy selects diverse batches of high-impact, in-domain sentences that result in a much more rapid improvement in translation performance than random and dissimilarity-based selection. This is reflected in objective indicators of translation quality (BLEU), and in terms of coverage as measured by the proportion of untranslated words in SMT hypotheses. We plan to evaluate the scalability of our approach by running simulations on a number of additional language pairs, domains, and corpus sizes.

An issue with iterative active learning in general is the cost of re-training the SMT system for each batch. Small batches provide for smooth performance trajectories and better error recovery at an increased computational cost. We are currently investigating incremental approaches that allow SMT models to be updated online with minimal performance loss compared to full re-training.

Finally, there is no inherent limitation in the proposed framework that ties it to a phrase-based SMT system. With suitable modifications to the input feature set, it can be adapted to work with various SMT architectures, including hierarchical and syntax-based systems.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 315, Morristown, NJ, USA. Association for Computational Linguistics.

Koby Crammer and Yoram Singer. 2001. Pranking with ranking. In *Advances in Neural Information Processing Systems 14*, pages 641–647. MIT Press.

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.

Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine Learning*, 28(2-3):133–168.

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.

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.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318, 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.