Active Learning in Example-Based Machine Translation

Rashmi Gangadharaiah
Carnegie Mellon University
Pittsburgh, PA
rgangadh@cs.cmu.edu

Ralf D. Brown
Carnegie Mellon University
Pittsburgh, PA
ralf@cs.cmu.edu

Jaime Carbonell
Carnegie Mellon University
Pittsburgh, PA
jgc@cs.cmu.edu

Abstract

In data-driven Machine Translation approaches, like Example-Based Machine Translation (EBMT) (Brown, 2000) and Statistical Machine Translation (Vogel et al., 2003), the quality of the translations produced depends on the amount of training data available. While more data is always useful, a large training corpus can slow down a machine translation system. We would like to selectively sample the huge corpus to obtain a sub-corpus of most informative sentence pairs that would lead to good quality translations. Reducing the amount of training data also enables one to easily port an MT system onto small devices that have less memory and storage capacity. In this paper, we propose using Active Learning strategies to sample the most informative sentence pairs. There has not been much progress in the application of active learning theory in machine translation due to the complexity of the translation models. We use a pool-based strategy to selectively sample instances from a parallel corpora which not only outperformed a random selector but also a previously used sampling strategy (Eck et al., 2005) in an EBMT framework (Brown, 2000) by about one BLEU point (Papineni et al., 2002).

1 Introduction

An EBMT system uses source-target sentence pairs present in a parallel corpus to translate new input source sentences. The input sentence to be translated is matched against the source sentences present in the corpus. When a match is found, the corresponding translation in the target language is obtained through sub-sentential alignment. The translation is generated from the partial target phrasal matches using a decoder. The motivation for using these systems is that they can quickly be adapted to new language pairs. EBMT systems in general have been found to require large amounts of data to function well and the quality of the target translations produced continues to improve as more and more data is added. However, many of the sentence pairs present in a parallel corpus do not contribute much to the translation quality. This could be due to the presence of poorly word-aligned sentence pairs, poorly translated sentences, spelling mistakes, repetition or redundancy in data. Using large amounts of data slows down the generation of the target sentence. In this paper, we use active learning to select useful sentence pairs from a large bilingual corpus.

Active Learning is a paradigm in Machine Learning, where a learner selects as few instances as possible (to be labelled by a labeller) and iteratively trains itself with the new examples selected. Supervised learning strategies require a large set of labeled instances to perform well. In many applications, unlabeled instances may be abundant but obtaining labels for these instances could be expensive and time-consuming. Active Learning was introduced to reduce the total cost of labeling.

The process of collecting the most useful examples for training an MT system is an active learning task, as a learner can be used to select these examples. This active learning strategy is not to be confused with translation model adaptation. In active learning, the assumption is that the test data is not available or known at selection time.

Different techniques exist for active learning (for a review see (Settles, 2009)), (i) membership query synthesis, (ii) stream-based selective sampling and (iii) pool-based active learning. Pool-based active learning is the most widely used technique. It assumes that there is a small set of labeled data and a large pool of unlabeled data. The

learner evaluates and ranks the unlabeled instances before selecting the best query.

There are a number of strategies a learner can follow to generate queries. In uncertainty sampling the learner queries instances that it is least certain how to label. In query-by-committee, multiple models are trained and the instance on which most models disagree is chosen as the query. Another strategy is to query the instance that would cause greatest change to the current model. Unfortunately these strategies are prone to outliers, which are common in MT systems. Instances can also be queried based on expected future error. This strategy is better resistant to outliers as it uses the unlabeled pool when estimating the future error. Density-weighted sampling strategy is also very common and is based on the idea that informative instances are those that are uncertain and representative of the input distribution. In this paper we will investigate these last two strategies.

Although active learning has been well studied in many natural language processing tasks, such as, Named-Entity Recognition (Shen et al., 2004), Parsing (Thompson et al., 1999), Word-sense disambiguation (Chen et al., 2006), not much work has been done in using these techniques to improve machine translation. (Eck et al., 2005) used a weighting scheme to select more informative sentences, wherein the importance is estimated using the unseen n-grams in the sentences that were previously selected. The length of the source sentence and actual frequency of the n-grams is used in their weighting scheme. Their experiments were based on the assumption that target sentences are not available at selection time, hence, no information from the target half of the data was used. Sentences were also weighted based on TF-IDF which is a widely used similarity measure in information retrieval. TF-IDF was used to find the most different sentence compared to the already selected sentences by giving it the highest importance i.e., the sentence with the lowest TF-IDF score is selected next. The TF-IDF approach did not show improvements over the other approach.

In (Eck et al., 2005) the system was evaluated against a weak baseline that selected sentences based on the original order of sentences in the training corpus. As adjacent sentences tend to be related in topic the number of new words added every iteration is low. A random selector would have been a stronger baseline. We show in this paper that random strategy would outperform (Eck et al., 2005) for EBMT systems. In this paper, we use a pool-based strategy that maximizes a measure of expected future improvement, to sample instances from a large parallel corpora. We also sample instances based on density of the input distribution and show that this modified sampling further improves the performance. Although the method is evaluated on a single language-pair in an EBMT paradigm, we expect to obtain improvements for other language pairs and other MT paradigms.

2 Description of the Method

Based on the properties of different active learning strategies (as described in the previous section), we conclude that a pool-based approach that selects sentence pairs based on expected future improvements is best suited for our EBMT task. The large corpus from which we select sentence pairs will be called the learner selector set, LSS. The sampled set with sentence pairs added so far into the active learning training set will be called the learner trained set, LTS. In a machine translation task there could be many possible ways to estimate the future improvement, such as translation BLEU scores. This would require retraining the MT system after adding each possible new sentence pair from the LSS into LTS, computing the BLEU score of the trained MT model on the remaining sentence pairs in LSS, and then adding the sentence pair which results in the best improvement in BLEU over the previous iteration to the LTS. Such a strategy is computationally infeasible, and so we suggest some modifications to the basic strategy that result in a substantial increase in speed without much loss in performance.

In this paper, we use a set of features that are much easier to compute than the BLEU score, noting that preliminary experiments indicated that they were good indicators of the sentence pair that would lead to best improvement in BLEU score over a test set. Also, to avoid having to estimate the improvement for every sentence pair in the LSS, we follow a cluster-then-sample approach that leads to a much smaller set (reduced set) that is still a representative of the LSS. We use a batch processing modification that speeds up the algorithm even further. We now describe the features used and the final score calculated from them.

$feature1$(Translation Score): Sentence pairs with high word alignment probabilities and new word

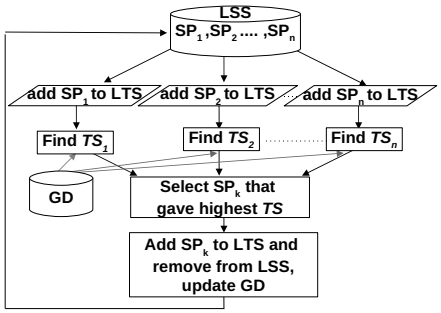

Figure 1: Active Learning strategy.

pair counts are arguably the most informative and are good candidates to improve translation quality. We start with a word-aligned bilingual corpus. In every iteration, a global dictionary (GD) which contains all word pairs added so far into LTS is consulted (Fig. 1). A scoring function is used to score each sentence pair (SP) in LSS and is defined as the sum of alignment scores in the reduced set of all those word pairs that are not present in GD but are present in the sentence pair. This score is then divided by the number of alignment scores that contributed to the summation. Normalizing the summation ensures that the word pairs added to GD are of high quality.

$feature2$(Alignment Score): the average of all word-alignment probabilities in the sentence pair.

The two features are linearly combined to obtain the totalscore(TS),

$$TS_{sentence\ pair} = \lambda_1 feature1 + \lambda_2 feature2$$

The sentence pair with the highest TS is added into the LTS and GD is updated with the new word pair entries found in the newly selected sentence pair. The feature values were normalized to have a mean of 0 and a variance of 1. For our preliminary experiments we gave equal weights to both the features ($\lambda_1 = \lambda_2 = 1$). To speed up the process further, a batch procedure is adopted. In every iteration, S sets with P points are randomly selected, where, S and P are parameters selected based on the amount of computation available. In this paper, $S = 100$ and $P = 10$. Each of these sets are scored using TS. The highest scoring set is added to the LTS in every iteration.

3 Experimental Setup and Preliminary results

A set of word-aligned 100k sentence pairs from FBIS English-Chinese data (NIST, 2003) was used

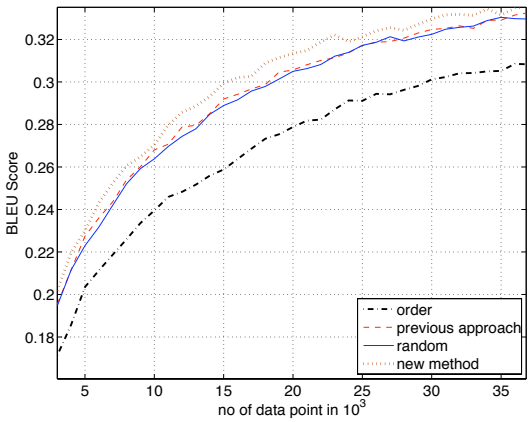

Figure 2: Comparison of (Eck et al., 2005), our method, random selection, selection based on the sentences in original order.

for the experiments. The reduced set was collected from 100k sentence pairs by first clustering the sentence pairs using the Lemur Clustering Application (Ogilvie and Callan, 2002) and picking sentence pairs randomly from each cluster such that it resembled the distribution of the entire word-aligned parallel corpus i.e., more sentence pairs were picked from denser regions and fewer from the less dense regions. The resulting set had 2056 sentence pairs. For the test set, 2500 sentence pairs were randomly chosen which had no overlap with the reduced set. To create the initial LTS, the remaining data was clustered using the Lemur Cluster Application and the centroid sentence-pairs were picked and ranked in the order of density with centroids from higher density regions appearing at the top. An initial LTS was formed by picking the first 2000 centroids. The remaining sentence pairs were used as the LSS.

3.1 Previous approach versus our method

We compared the method suggested in (Eck et al., 2005) with two baselines tested on the test set, one in which the sentence pairs were selected based on the original order of the sentences in the corpus, the other with sentence pairs randomly selected. The same LTS was used for (Eck et al., 2005). From Fig. 2, it can be seen that (Eck et al., 2005) outperforms the first baseline but there is no clear improvement over the second random baseline.

Our method was also compared with the same two baselines. From Fig 2, it can be seen that our method outperforms both baselines and (Eck et al., 2005) by 1 BLEU point. All the approaches were

run until the LTS contained 65,000 sentence pairs, the plot in Fig 2 shows BLEU scores only up to 37,000 sentence pairs as after this point the scores for the approach (Eck et al., 2005), our method and random had no significant difference.

3.2 Incorporating Density Information

Density weighted sampling performs uncertainty or query-by-committee sampling from dense regions. Since density weighted sampling strategies sample points from maximal-density regions, they help in forming the initial decision boundary where it affects the most remaining unsampled points. Density-based sampling methods are known to perform well in the initial iterations when the amount of data in LTS is small. We performed an initial experiment to see if this was true even in MT. For our preliminary experiments, we only sampled sentence pairs from the dense regions and did not use uncertainty or query-by-committee strategies. What we aim to sample here are the centroids which we believe are a good representation of dense regions. For this, the LSS was first clustered using the Lemur Cluster Application. In an iteration, P centroids from the most dense regions were sampled and their performance was tested on the test set (Fig 3). In the next iteration, P centroids from the next most P dense regions were picked. This process was iteratively performed until there were no clusters (with more than 3 sentence pairs) left. This took roughly 800 iterations to exhaust the centroids. For the remaining iterations, the method explained in Fig 1 was applied. As predicted, from Fig 3, it can be seen that this method performs better than the approach in Fig 1 up to 11,000 sentence pairs but its performance drops when more data is added to the LTS using the approach in Fig 1.

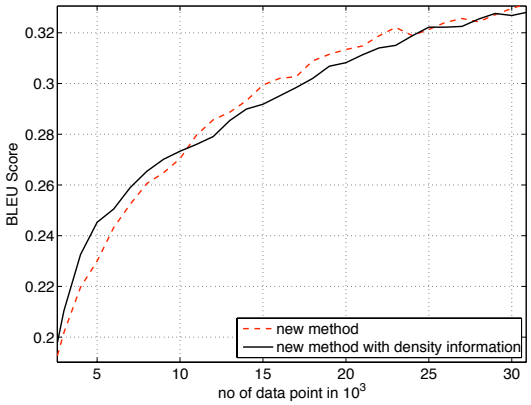

Figure 3: Comparison of the density-based method and the non-density method in Fig 1.

4 Conclusion and Future work

In this paper, we used a pool-based strategy to selectively sample instances from a word aligned parallel corpora which not only outperformed a random selector but also a previously suggested sampling strategy in an EBMT framework. As future work, we would like to perform experiments with different sizes of initial LTS and larger sizes of LSS where we expect to see more improvements. In our batch processing framework, we sampled S sets each of size P randomly, it would be interesting to see the performance when we use density or uncertainty strategies to pick samples. We also used a density-based sampling strategy which was found to help only in the initial iterations, and as future work, we would like to combine it with other sampling strategies.

References

B. Settles. 2009. Active Learning Literature Survey. *Computer Sciences Technical Report 1648,* University of Wisconsin-Madison.

C. A. Thompson, M. E. Califf, R. J. Mooney 1999. Active Learning for Natural Language Parsing and Information Extraction. *Proc. of ICML.*

D. Shen, J. Zhang, J. Su, G. Zhou, C. L. Tan. 2004. Multi-criteria-based active learning for named entity recognition. *Proc. of the 42nd Annual Meeting of the ACL.*

J. Chen, A. Schein, L. Ungar, M. Palmer. 2006. An empirical study of the behavior of active learning for word sense disambiguation. *Proc. of HLT/NAACL.*

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proc. of the 20th Annual Meeting of the ACL.*

M. Eck, S. Vogel, and A. Waibel. 2005. Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF. *Proc. of IWSLT.*

P. Ogilvie and J. Callan 2002. Experiments using the Lemur toolkit. *Proc. of TREC 2001,* NIST, special publication 500-250.

R. D. Brown 2000. Example-Based Machine Translation at Carnegie Mellon University. *The ELRA Newsletter, vol 5:1.*

S. Vogel, Y. Zhang, A. Tribble, F. Huang, A.Venugopal, B. Zhao, and A. Waibel. 2003. The CMU Statistical Translation System. *Proc. of MT Summit IX.*