Multi-Strategy Approaches to Active Learning for Statistical Machine Translation

Vamshi Ambati, Stephan Vogel and Jaime Carbonell
{vamshi,vogel,jgc}@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213, USA

Abstract

This paper investigates active learning to improve statistical machine translation (SMT) for low-resource language pairs, i.e., when there is very little pre-existing parallel text. Since generating additional parallel text to train SMT may be costly, active sampling selects the sentences from a monolingual corpus which if translated would have maximal positive impact in training SMT models. We investigate different strategies such as density and diversity preferences as well as multi-strategy methods such as modified version of DUAL and our new ensemble approach GraDUAL. These result in significant BLEU-score improvements over strong baselines when parallel training data is scarce.

1 Introduction

Large scale parallel data generation for new language pairs requires intensive human effort and availability of experts. Only a few languages in the world enjoy sustained research interest and continuous financial support for development of automatic translation systems. For most remaining languages there is very small interest or funding available. Therefore it becomes immensely difficult and costly to provide Statistical Machine Translation (SMT) systems for such languages.

In this paper we resort to Active Learning (AL) techniques for building MT systems for minority languages. In active learning, the learner has access to a large pool of unlabeled data and sometimes a small portion of seed labeled data. Therefore the objective of an active learner is to rank a set of instances in an optimal manner for an external oracle to label them so as to provide maximal benefit to the learner. Annotations in MT can be of various kinds depending upon the paradigm of translation. In our case since we work with a Statistical Machine Translation system (SMT), our annotations are to seek target-language translation for a source-language sentence. We propose a novel query strategy, Density Weighted Diversity Sampling (DWDS) which focuses on both diversity and density metrics in selecting a sentence. Our approach works significantly better than other baselines, as reported in our experiments section.

We also explored multiple active learning query strategies for the task of sentence selection. We observe that some methods perform well in initial phases where very few instances have been sampled, while others perform better in later operating ranges upon substantial sampling. For instance, density estimation methods (Nguyen and Smeulders, 2004) perform well with minimal labeled data since they sample from maximal-density unlabeled regions, and thus build an MT model that is capable of translating majority of the remaining unlabeled data. On the other hand, diversity sampling method focuses more on improving recall by favoring unseen words irrespective of their representativeness in the data. With awareness of the performance of a query strategy under a particular operating range we propose multi-strategy query methods that can do a better job of performing well under a larger operating range by selecting optimal query strategy for different operating ranges.

We consider two different strategies for sentence selection in MT, that have varying returns in different phases of translation. The first method is our density oriented approach (DWDS), which focuses on maximally-dense n-grams in the unlabeled data. The second method is a Diversity sampling (DIV) approach which focuses on n-grams that are different from those already present in the labeled data. Inspired by the work in (Donmez et al., 2007), we propose a multi-strategy approach (DUAL) to switching from a DWDS to a DIV strategy. While (Donmez et al., 2007) switch from a density focused to an uncertainty focused strategy, we use a diversity focused approach. Uncertainty of a model has been used as a successful active learning strategy (Lewis and Catlett, 1994). For the task of translation, we choose diversity as a strategy instead of 'uncertainty' as our experiments show that diversity is much faster to compute and the performance is very similar to uncertainty sampling approach. Computing uncertainty of a statistical translation model requires retraining of the model across iterations, which is time consuming. We also extend the DUAL approach and propose a novel ensemble approach called GraDUAL. While DUAL estimates a switch over point to transit to a second querying strategy, GraDUAL chooses an operating range in which it performs a gradual switch over. In the switch over range, we perform a dynamically weighted interpolation for sampling under the two approaches in consideration. This ensures a smooth transition from one strategy to the other and is robust to noise that may false project one query strategy to be better than the other.

The rest of the paper is organized as follows. In section 3 we present our framework for active learning in SMT and discuss our sentence selection algorithm. Section 4 describes DUAL, a multi-strategy approach that focuses on switching between two strategies. Section 5 discusses GraDUAL, another hybrid approach that addresses some of the issues with DUAL. Section 6 presents experiments and results on Spanish-English language pair. We conclude with discussion of related work in Section 2.

2 Related Work

Active learning has been applied to various fields of Natural Language Processing like statistical parsing, entity recognition among others (Hwa, 2004; Steedman et al., 2003; Shen et al., 2004). In case of MT, the potential of active learning has remained largely unexplored. For SMT, application of active learning has been focused on the task of selecting the most informative sentences to train the model, in order to reduce cost of data acquisition. Recent work in this area discussed multiple query selection strategies for a Statistical Phrase Based Translation system (Haffari et al., 2009). Their framework requires source text to be translated by the system and the translated data is used in a self-training setting to train MT models. (Gangadharaiah et al., 2009) use a pool-based strategy that maximizes a measure of expected future improvement, to sample instances from a large parallel corpus. Their goal is to select the most informative sentence pairs to build an MT system, and hence they assume the existence of target-side translations along with the source-side sentences. We however are interested in selecting most informative sentences to reduce the effort and cost involved in translation.

(Eck et al., 2005) use a weighting scheme to select more informative sentences, wherein the importance is estimated using unseen n-grams in previously selected sentences. Although our selection strategy has a density based motivation similar to theirs, we augment this by adding a diminishing effect to discourage the domination of density and favor unseen n-grams. Our approach, therefore, naturally works well in pool-based active learning strategy when compared to (Eck et al., 2005). In case of instance-based active learning, both approaches work comparably, with our approach working slightly better.

Ensemble approaches have been proposed in active learning literature and have been successfully applied to classification tasks (Melville and Mooney, 2004; Freund et al., 1997). Trading off between density and uncertainty has been the focus of several of these active learning strategies (McCallum and Nigam, 1998; Nguyen and Smeulders, 2004). (Baram et al., 2004) propose an online algorithm to select among multiple strategies and decide the strategy to be used for each iteration. Most notably our

approach is inspired from the DUAL approach proposed in (Donmez et al., 2007), where the authors differ from earlier ensemble approaches by not focusing on selecting the best strategy for the entire task, but by switching between multiple strategies over different ranges. Ensemble methods for active learning in MT have not been explored to our knowledge. (Haffari et al., 2009) address an interesting technique of combining multiple query strategies for the task of sentence selection. Tuning the weights of combination and optimizing towards translation quality is computationally expensive, and their approach does not perform better than the best performing single strategy approach.

3 Active Learning for Machine Translation

We now discuss our general framework for active learning in SMT and then discuss the sentence selection approach we use to pick informative sentences.

3.1 Active Learning Setup

We start with an unlabeled dataset $U_0 = \{f_j\}$ and a seed labeled dataset $L_0 = \{(f_j, e_j)\}$, where labels are translations. We then score all the sentences in the U_0 according to our selection strategy and retrieve the best scoring sentence or a small batch of sentences. This sentence is translated and the sentence pair is added to the labeled set L_0. However, re-training and re-tuning an SMT system after translating every single sentence is computationally inefficient and may not have a significant effect on the underlying models. We, therefore continue to select a batch of N sentences before retraining the system on newly created labeled set $L_{k=1}$. Our framework for active learning in SMT is discussed in Algorithm 1.

Algorithm 1 ACTIVE LEARNING FOR SMT

```
Given Labeled Data Set : L_0
Given Unlabeled Data Set: U_0
for k = 0 to T do
    for i = 0 to N do
        s_i = Query(U_i,L_i)
        t_i = Human Translation for s_i
        S_k = S_k ∪ (s_i, t_i)
    end for
    U_{k+1} = U_k - S_k
    L_{k+1} = L_k ∪ S_k
    Re-train MT system on L_{k+1}
end for
```

3.2 Sentence Selection

Our sentence selection strategy to be independent of the underlying SMT system or the models and has been shown to perform well (Ambati et al., 2010). For the sake of comprehensiveness we discuss the approach here as well. We use only monolingual data U and bilingual corpus L to select sentences. This makes our approach applicable to any corpus-based MT paradigm and system, even though we test with SMT. The basic units of an SMT system are phrases and therefore we measure the informativeness of a sentence in terms of the consisting phrases . Our scoring strategy is shown in equation below. We select sentences that have the most representative n-grams and have not yet been seen in the bilingual corpus. Representativeness or the 'density' of a sentence is computed by $P(x|U)$ as relative likelihood estimates of the n-gram x in the unlabeled monolingual data. We also introduce a decay on the density of an n-gram based on its frequency in the labeled data. Novelty or 'uncertainty' is computed as the number of new phrases that a sentence has to offer. We compute the final score of a sentence as the harmonic mean of both these metrics with a tunable parameter 'β', that helps us balance the novelty and density factors. We choose $\beta = 1$ and $\lambda = 1$ for our current experiments. Thus far we have only considered n-grams of size upto 3.

$$
\begin{aligned}
d(s) &= \frac{\sum_x^{Phrases(s)} P(x|U) * e^{-\lambda count(x|L)}}{|Phrases(S)|} \\
u(s) &= \frac{\sum_x^{Phrases(s)} \alpha}{|Phrases(s)|} \\
\alpha &= \begin{cases} 1 & x \notin Phrases(L) \\ 0 & \end{cases} \\
Score(s) &= \frac{(1+\beta^2) d(s) * u(s)}{\beta^2 d(s) + u(s)}
\end{aligned}
$$

4 DUAL Strategy

Let us consider the DWDS approach in more detail. It has two components for scoring a sentence

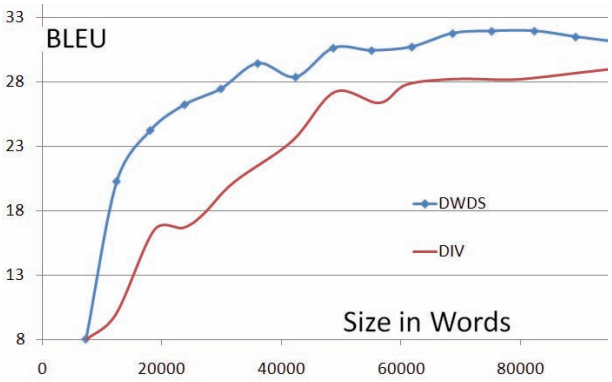

Figure 1: Density vs Diversity performance curves

S, a density component $d(s)$ and a diversity component $u(s)$ as mentioned in the previous section. The DWDS approach favors those sentences that contain dense n-grams and thus has the largest contribution to the improvement of translation quality. Combining diversity with density of the underlying data is a well known ensemble technique in active learning that improves performance (Nguyen and Smeulders, 2004). Now consider DIV selection criteria that favors sentences with unseen n-grams. Such a method is prone to selecting uncommon sentences that add very little information to the translation model.

Figure 1 displays the translation performance of 'DWDS' and 'DIV' on a held-out dataset as measured in BLEU vs size of labeled training data in words. One observation is that DWDS, after rapid initial gains exhibits very slow incremental improvements. Diversity sampling shows continuous and consistent improvements over a longer operating range. We computed overlap of instances selected by the two methods and found that there is a very low overlap, showing that there is significant disagreement in sentence selection by the two approaches.

In the initial phases of evolution of an MT system, there is very little or no labeled data, hence every sentence is highly diverse. DWDS can pick high density sentences which may have been scored lower by the DIV technique. As more data is labeled, explicitly dense sentences may not be found anymore. Therefore, DWDS may score sentences with moderate density higher than the sentences with high diversity, there by making this criterion suboptimal. It is this weakness that we would like to address using the DUAL approach.

DUAL approach has been applied successfully for text classification problems in (Donmez et al., 2007). We adapt this approach to the task of MT. DUAL approach performs sentence selection using DWDS until a certain switching point is reached. A switching point is that point in the learning process, beyond which DWDS approach tends to provide only slow improvements. In other words, at a switching point we observe the density component of DWDS dominating the diversity component. Beyond the switching point, we use DIV active learning strategy for sentence selection. Algorithm 3 provides details of the DUAL approach.

Algorithm 2 ITERATION

Given Unlabeled Data Set: U_k
Given Labeled Data Set : L_k
for $i = 0$ to N **do**
    s_i = Query(U_i,L_i)
    t_i = Human Translation for s_i
    $S_k = S_k \cup (s_i, t_i)$
end for
$U_{k+1} = U_k$ - S_k
$L_{k+1} = L_k \cup S_k$
Re-train MT system on L_{k+1}

Algorithm 3 DUAL APPROACH

Given Unlabeled Data Set: U_0
Given Labeled Data Set : L_0
$k = 0$
$SWITCH = false$
while $SWITCH = false$ **do**
    Query = DWDS
    ITERATION(U_i,L_i)
    β = Compute TTR(U_k,L_k)
    if $\beta > \delta$ **then**
        $k = k + 1$
        SWITCH = $true$
    end if
end while
for $k = k$ to T **do**
    Query = DIVERSITY
    ITERATION(U_i,L_i)
end for

4.1 Switching Point

Estimation of the switching point is the key to the success of DUAL approach. Switching too early may take away the benefit of DWDS approach, and switching too late may not yield the benefits of DIV sampling approach.

Let us first consider an ideal scenario for switching where we have access to the learning curves from DWDS and DIV, like the ones shown in Figure 1. Looking at the curves , one natural choice for a switching point is where the slope of DWDS learning curve drops lower than the DIV learning curve. As our experiments later show, this switching point does in fact perform well in terms of translation quality.

The problem with this above approach is that it assumes availability of both the curves that have been produced independently. A learning curve here is over the number of translations on x-axis and the direct improvement in translation quality on y-axis as measured by BLEU metric for MT evaluation. In order to compute such a curve we need to select a batch of sentences using a querying strategy, translate the batch (or a subset), retrain and retest on held out dataset to observe the gradient of improvement across iterations. This is not feasible as we will be spending twice the amount of cost and also retrain the MT system twice. Although computation is not an issue, doubling the cost is unacceptable. Hence, we would like to identify the switching point by an approximation of the translation improvement, which is easy to compute.

We propose a surrogate metric based on types and token ratios that are computed only using source sentences of the labeled data. Type vs token curves indicate growth of vocabulary of the corpus. We use such curves to understand the effects of 'Density' and 'Diversity' in active learning based sentence selection. Figure 2 shows such a curve on the Spanish-English dataset. Density based approaches place an emphasis on the distribution of the data , and therefore provide a larger coverage for tokens. At the same time, the diversity focused component ensures aggregation of new types.

We propose a metric called 'type-token ratio'(TTR) that highlights the balance between the tokens and types of the unlabeled data, which are represented in a corpus selected using an active learning querying method. We can compute such a metric as shown below.

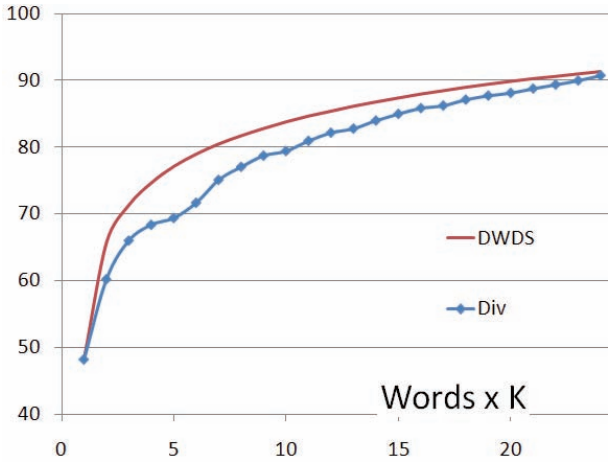

Figure 2: TTR curves for Spanish-English

$$
\begin{aligned}
Typ_k(L_k, U_0) &= \frac{\sum_x^{Phrases(U_0)} \alpha}{\|Phrases(U_0)\|} \\
\alpha &= \begin{cases} 1 & x \in Phrases(L_k) \\ 0 & \end{cases} \\
Tok_k(L_k, U_0) &= \frac{\|Phrases(L_k) \cap Phrases(U_0)\|}{\|Phrases(U_0)\|} \\
TTR_k(L_k, U_0) &= \frac{2 * Typ_k * Tok_k}{(Typ_k + Tok_k)}
\end{aligned}
$$

It is inexpensive to compute TTR curves for both the DWDS and DIV query methods. The switching point is chosen where the slope of DWDS curve is lower than the DIV curve by a margin, shown as a constraint below. We set δ to be a very small number, 0.02 in our experiments.

$$\Delta(DWDS_k) > \Delta(DIV_k) + \delta \qquad (1)$$

5 GraDUAL Approach

One of the problem with DUAL is that the success of the method depends largely on the robust estimation of the switching point. This is not robust in cases where noise may cause a temporary dip in the slope of the TTR curve for DWDS. Noise can cause a false switching from one strategy to another, even when it is not the right sampling strategy to be exploited.

Given the multiple factors and parameters in training an MT system, it is natural to expect such instable behavior in the initial phases of the system. We therefore propose a different hybrid strategy called 'GraDUAL', which gradually switches from DWDS to DIV strategies. We do not assume the existence of a 'switching point', but try to estimate a 'switching range' during which the transition between strategies takes place.

GraDUAL approach, as described in 4, is motivated from the concept of 'exploration vs. exploitation'. This approach exploits the sampling strategy that is evidently better in a given range. We compute the slope of the TTR curve between two consecutive iterations as Δ. A positive and increasing slope indicates good performance of the approach. When comparing two different TTR curves we will have operating ranges where the slopes do not project a clear winner. In such cases, GraDUAL approach suggests sampling from both strategies, with a gradual shift towards the second technique. The rate of shift is controlled by the parameter $f(\beta)$. In our current work we use a constant $f(\beta) = 0.8$ to sample 80% from the best performing strategy and 20% from the second. We are experimenting with other functions for $f(\beta)$.

$$
\begin{aligned}
\beta &= Abs(\Delta(DWDS) - \Delta(DIV)) \\
\alpha &= \begin{cases} 1 & \beta > \delta \\ 0 & \beta < \delta \\ f(\beta) & \end{cases} \\
Score(s) &= \alpha DWDS(s) + (1-\alpha)DIV(s)
\end{aligned}
$$

Algorithm 4 GRADUAL APPROACH

1: Given Labeled Data Set : L_0
2: Given Unlabeled Data Set: U_0
3: β = 1
4: **for** $k = 0$ to T **do**
5: Query method = GraDUAL
6: β = Compute Ratio(U_k,L_k)
7: ITERATION(U_i,L_i,β)
8: **end for**

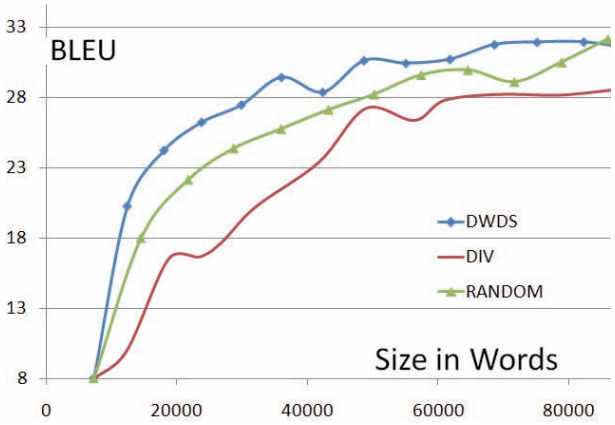

Figure 3: Spanish-English results:Single-Strategy

6 Experiments

6.1 Setup

We perform our experiments on the Spanish-English language pair in order to simulate a resource-poor language pair. We have parallel corpora and evaluation data sets for the Spanish-English language pair allowing us to run multiple experiments efficiently. We use BTEC parallel corpus (Takezawa et al., 2002) from the IWSLT tasks with 127K sentence pairs. We use the standard Moses pipeline (Koehn et al., 2007) for extraction, training and tuning our system. We built an SRILM language model using English data consisting of 1.6M words. While experimenting with data sets of varying size, we do not vary the language model. The weights of the different translation features are tuned using standard MERT (Och, 2003). Our development set consists of 506 sentences and test set consists of 343 sentences. We report results on the test set.

6.2 Results: AL for MT

We first test the performance of our active learning sentence selection strategy. We start with an initial system trained on 1000 sentence pairs. We then train the system iteratively on datasets of increasing size. In each iteration, we first selectively sample 1000 Spanish sentences from source side of the entire corpus. We simulate human translation in our experiment, as we already have access to the translations from the BTEC corpus. We then retrain, tune and test the system to complete the iteration.

We compare our results with two strong baselines.

First is a random baseline, where sentence pairs are sampled at random from the unlabeled dataset. Random baselines are strong as they tend to simulate the underlying data distribution when sampled in large numbers. The second baseline is where we select data based on the order it appears in BTEC corpus. As seen in Figure 3, our active learning strategy performs better than the two baselines. The x-axis in the graph is the number of words of parallel data used for training the system, and y-axis shows performance as measured by BLEU on a held out dataset. One way to read the results is that for the same amount of parallel sentences used, active learning helps to select more informative sentences and hence achieves better performance. Alternatively, we can understand this as follows. Given an MT system, active learning strategy uses less number of sentences to reach a desired accuracy, thereby reducing cost of acquiring data. For the same amount of data trained on, DWDS selection strategy achieves 2 BLEU points higher than random selection strategy. Another observation from the curve is that DWDS achieves 30.5 BLEU points on a held out test set by training on about 27% less data than random selection.

6.3 Results: Hybrid AL approaches

We evaluate our multi-strategy approaches and present results. We first compare the robustness of our surrogate metric based switching strategy with 'oracle switching'. Figure 4 shows results on development set when switching using BLEU score based learning curves. An oracle then visually inspects and selects an iteration to switch where the DWDS learning curve's slope is lower than that of the DIV learning curve by a margin. We compare this with results from executing the DUAL approach using our surrogate metric, TTR, to decide the switching point. We observe that switching using feedback from TTR works on par with BLEU, and is also easier to compute. We therefore report experiments with multi-strategy approaches using the TTR surrogate.

In Figure 5 we compare DUAL and GraDUAL approaches to our best performing active learning strategy DWDS and also DIV. When considered independently, AL approaches have a disadvantage that they are hindered by the selection of data made in the earlier iterations. The point of switching strategies is that the second strategy can build on top of better selections made by its predecessor. The results show a similar trend. We observe that both our multi-strategy approaches that include DIV switching strategy perform significantly better than the two baseline approaches even when DIV does not do better than DWDS in isolation. From the results it can also be seen that although GraDUAL and DUAL perform comparably, GraDUAL displays a smoother transition from one strategy to the other. Overall, using multi-strategy ensemble approaches we have shown that MT systems can reach significantly better performance while requiring much lower amounts of data. At different points on the curves, prior to convergence, we have performed bootstrapped sampling based significance tests with the baseline and see that the p-value varies between 0.02 and 0.11 (averaging at 0.06). So the reported results are statistically indicative and with a larger experiment (more observations) should prove statistically significant.

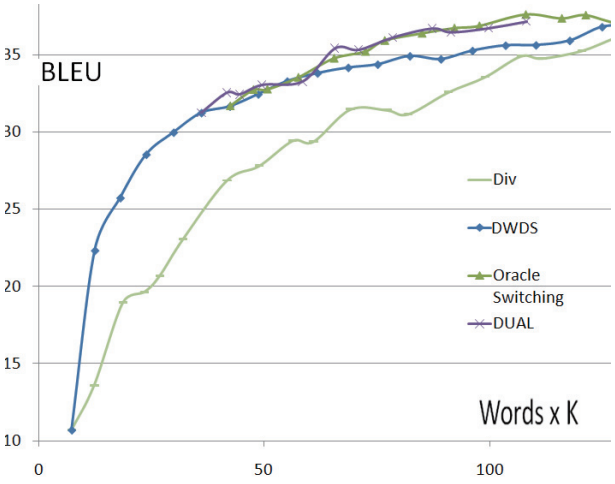

Figure 4: BLEU vs Surrogate comparison on Devset

7 Conclusion and Future Work

In this paper we have addressed three issues. We proposed active learning sentence selection strategy for Statistical Machine Translation whose performance is comparable to the state-of-art approaches. We improved our best performing AL strategy by a modified version of the DUAL (Donmez et al., 2007) approach and a novel and robust GraDUAL approach. We experimented our approaches on

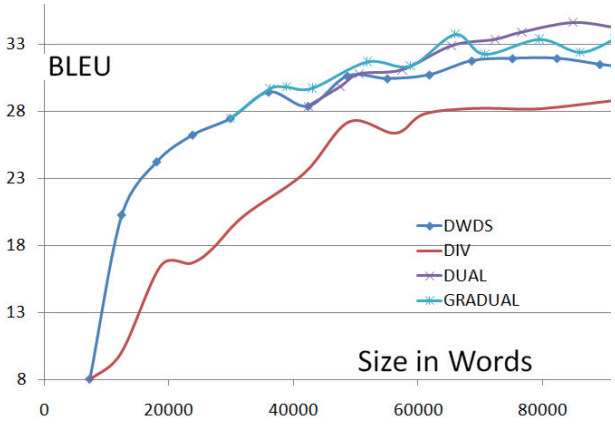

Figure 5: Spanish-English results: Multi-Strategy

Spanish-English language pair and have shown significant improvements. In future we would like to improve on our surrogate based 'switching point' identification. We would also like to deploy our active learning approaches in developing MT systems for other language pairs and domains. Active learning for MT has not yet been explored in its full potential. Much of the literature and this paper, have explored one task - selecting sentences to translate and add to the training corpus. We would like to explore other tasks that have value in context of an SMT system such as, translation of individual words or phrases, hand-generation of word alignments, corrections of system-generated translations. We would also like to apply our ensemble query selection strategies to these tasks as well.

References

Ambati, Vamshi, Stephan Vogel, and Jaime Carbonell. 2010. Active learning and crowd-sourcing for machine translation. In *Proceedings of the LREC 2010*.

Baram, Yoram, Ran El-Yaniv, and Kobi Luz. 2004. Online choice of active learning algorithms. *J. Mach. Learn. Res.*, 5:255–291.

Donmez, Pinar, Jaime G. Carbonell, and Paul N. Bennett. 2007. Dual strategy active learning. In *ECML*, pages 116–127.

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2005. Low cost portability for statistical machine translation based on n-gram coverage. In *Proceedings of MT Summit X*, Phuket, Thailand.

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

Gangadharaiah, Rashmi, Ralf Brown, and Jaime Carbonell. 2009. Active learning in example-based machine translation. In *Proc. of the 17th Nordic Conf of Computational Linguistics NODALIDA 2009*.

Haffari, Gholamreza, Maxim Roy, and Anoop Sarkar. 2009. Active learning for statistical phrase-based machine translation. In *Proceedings of HLT NAACL 2009*, pages 415–423, Boulder, Colorado, June. Association for Computational Linguistics.

Hwa, Rebecca. 2004. Sample selection for statistical parsing. *Comput. Linguist.*, 30(3):253–276.

Koehn, Philipp, Hieu Hoang, Alexandra Birch Mayne, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL Demonstration Session*.

Lewis, David D. and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. In *In Proc. of the Eleventh International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann.

McCallum, Andrew and Kamal Nigam. 1998. Employing em and pool-based active learning for text classification. In *ICML*, pages 350–358.

Melville, Prem and Raymond J. Mooney. 2004. Diverse ensembles for active learning. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 74, New York, NY, USA. ACM.

Nguyen, Hieu T. and Arnold Smeulders. 2004. Active learning using pre-clustering. In *ICML*.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the Association for Computational Linguistics*, Sapporo, Japan, July 6-7.

Shen, Dan, 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 ACL*, page 589, Morristown, NJ, USA. ACL.

Steedman, Mark, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Example selection for bootstrapping statistical parsers. In *NAACL '03: Proceedings of the 2003 NAACL*, pages 157–164, Morristown, NJ, USA. ACL.

Takezawa, Toshiyuki, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Towards a broad-coverage bilingual corpus for speech translation of travel conversation in the real world. In *Proceedings of LREC 2002, Las Palmas, Spain*.