Active Learning and Crowd-Sourcing for Machine Translation

Vamshi Ambati, Stephan Vogel, Jaime Carbonell

Language Technologies Institute
Carnegie Mellon University, Pittsburgh, PA 15213, USA
{vamshi,vogel,jgc}@cs.cmu.edu

Abstract

In recent years, corpus based approaches to machine translation have become predominant, with Statistical Machine Translation (SMT) being the most actively progressing area. Success of these approaches depends on the availability of parallel corpora. In this paper we propose Active Crowd Translation (ACT), a new paradigm where active learning and crowd-sourcing come together to enable automatic translation for low-resource language pairs. Active learning aims at reducing cost of label acquisition by prioritizing the most informative data for annotation, while crowd-sourcing reduces cost by using the power of the crowds to make do for the lack of expensive language experts. We experiment and compare our active learning strategies with strong baselines and see significant improvements in translation quality. Similarly, our experiments with crowd-sourcing on Mechanical Turk have shown that it is possible to create parallel corpora using non-experts and with sufficient quality assurance, a translation system that is trained using this corpus approaches expert quality.

1. Introduction

Corpus based approaches to automatic translation like Example Based and Statistical Machine Translation systems use large amounts of parallel data created by humans to train mathematical models for automatic language translation (Koehn et al., 2003). Large scale parallel data generation for new language pairs requires intensive human effort and availability of experts. It becomes immensely difficult and costly to provide Statistical Machine Translation (SMT) systems for most languages due to the paucity of expert translators to provide parallel data. Even if experts are present, it appears infeasible due to the impending costs. Most research in Machine Translation (MT) has been made possible by massive data made available through a small number of expensive data entry efforts. European parliament proceedings, Canadian Hansards, BTEC (Takezawa et al., 2002) are examples of a few initiatives that have made research in Statistical Machine Translation attractive for some language-pairs. Given that there is not sufficient interest for large audiences in most remaining language pairs, MT systems typically remain unbuilt for foreseeable future.

With the advent of online market places such as Amazon Mechanical Turk [1], it is now easier to reach bilinguals on the Web than ever before, even if most of them are not expert translators. Researchers in the Natural Language Processing community are quickly exploiting 'crowd-sourcing' for acquisition of labeled data (Snow et al., 2008), where annotation tasks are farmed out to a large group of users on the web utilizing micro payments.

In the case of machine translation, although services like Mechanical Turk (MTurk) have opened doors to tap human potential, they do not guarantee translation expertise nor large-volume availability of translators. We propose Active Crowd Translation (ACT), a framework for tying active learning with crowd-sourcing to reduce costs and make best use of human resources for generating the most useful parallel data for building machine translation systems. Active learning approaches help us identify sentences, which if translated have the potential to provide maximal improvement to an existing system. Crowd-sourcing techniques, on the other hand help us reach a significant number of translators at very low costs. This is very apt in a minority language scenario, where cost plays a major role. This paper addresses the following contributions:

- We propose and implement an end-to-end human-in-the-loop translation system framework called Active Crowd Translation, that combines online, non-expert, human translators with an automatic MT system.
- We propose active learning strategies for the problem of 'sentence selection' to reduce cost of building parallel data for MT systems.
- We propose approaches to the problem of 'translation selection', i.e how to select a good translation from among the multiple translations provided by several non-experts.

The rest of the paper is organized as follows. In section 2., we discuss state-of-the-art and relevant work in the areas of active learning and crowd-sourcing. Section 3. describes our ACT paradigm and the implementation framework. We provide a detailed discussion crowd-sourcing using Amazon Mechanical Turk and discuss some of the challenges in section 4.. In Section 5. we discuss active learning for the task of sentence selection. Section 6. highlights the issues of quality in crowd data and our approach of translation selection. Our experimental setup and results are presented in section 7..

2. Related Work

2.1. Active Learning

In active learning, a few labeled instances are typically provided together with a large set of unlabeled instances. The objective is to rank a set of instances in an optimal way for an external oracle to label them. The underlying system is then re-run to improve performance. This continues in

[1] http://www.mturk.com/mturk/

an iterative fashion for convergence - which typically is a threshold on the achievable performance before exhausting all the unlabeled data set.

Active learning has been applied to Statistical Parsing (Hwa, 2004; Tang et al., 2001) to improve sample selection for manual annotation. In case of MT, active learning has remained largely unexplored. Some attempts include training multiple statistical MT systems on varying amounts of data, and exploring a committee based selection for re-ranking the data to be translated and included for retraining (Callison-burch, 2003). But this does not apply to training in a low-resource scenario where data is scarce.

Recent work 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 corpora. Their goal is to select the most informative sentence pairs to build an MT system, and hence they assume the existence of target sides translations along with the source sides. 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. (Callison-burch, 2003) discuss approaches for applying active learning to machine translation, but there has not been an implementation or deployment of the algorithms in it.

2.2. Crowd-sourcing

(Snow et al., 2008) discuss usability of annotations created by using Mechanical Turk for a variety of NLP tasks, primarily supervised learning tasks for classification. These tasks included word sense disambiguation, word similarity, textual entailment, and temporal ordering of events. (Hsueh et al., 2009) perform feasibility analysis of using crowd data for sentiment analysis. Recent efforts in MT include feasibility studies for using crowd sourcing techniques for MT Evaluation, where users are provided with translations from multiple systems and asked to select the correct one (Callison-Burch, 2009; Zaidan and Callison-Burch, 2009).

3. ACT: Active Crowd Translation

In the currently predominant data-driven paradigm to machine translation, an expert is provided with a defined set of source language sentences which are then translated into the target language. Such data is then used to train an MT system. In an Active Crowd Translation (ACT) framework, the key idea is a 'crowd' of non-experts and experts actively participating in the translation of sentences as deemed useful by an active learning module. As seen in Figure 1, the ACT module first selects sentences from a monolingual source corpus, which it thinks are the most informative sentences to be translated next. We assume that the monolingual data is from a similar domain as the unseen test corpus. An interesting challenge is to relax the assumption of similar domain. The sentences are then posted to a crowd-sourcing online marketplace like Mechanical Turk, where multiple translators can translate a sentence. Translations from the crowd are then compared with each other and possibly with external sources for quality assurance. Best fit translations are then selected from the multiple translations. The MT system is re-trained with this new parallel data. Feedback from the MT system then drives further iterations of the active data collection.

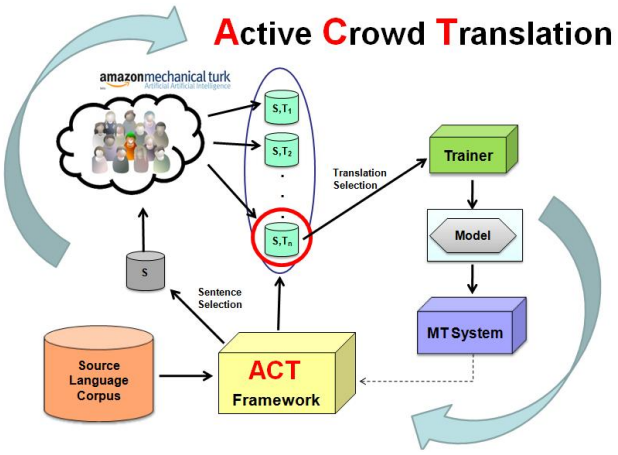

Figure 1: ACT Framework

4. Crowd-sourcing via Mechanical Turk

We use Amazon Mechanical Turk (MTurk) as our crowd-sourcing platform. Mturk is an online marketplace that enables programmers or users with data requirement to co-ordinate with humans in order to create it, via crowd-sourcing. 'Requesters', who are typically in need of human intervention for labeled data, can post annotation tasks known as HITs (Human Intelligence Tasks). These tasks that are typically easier and quicker for humans to complete than for machines, language translation being a striking example. It is still difficult to produce human quality translations for systems. Workers, also known as turkers, can then browse among existing tasks and complete them for a payment provided by the Requester. In case of machine translation a HIT is a task where a turker is provided with one or more sentences in the source language to be translated to a target language.

4.1. Expert vs. Non-Experts

Quality assurance is always a concern with an online crowd that has a mixture of experts and non-experts. Amazon provides a couple of mechanisms to help enforce preliminary quality assurance. A requester can require that all workers meet a particular set of qualifications, such as sufficient accuracy on a small test set or a minimum percentage of previously accepted submissions. One can also place location

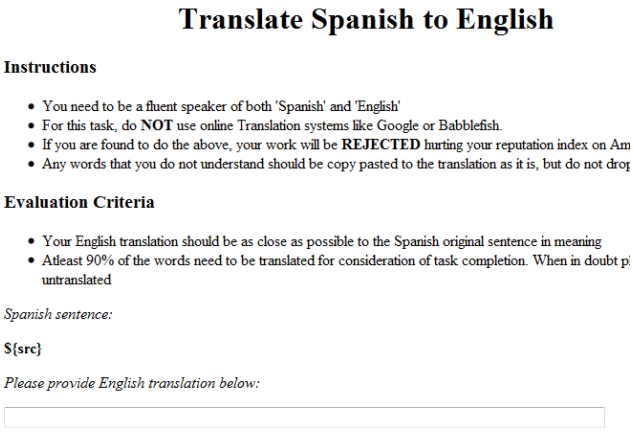

Figure 2: Sample HIT template on MTurk

based restrictions, which in our case can help us seek translations from a particular language-speaking country. For example, we can place a restriction of selecting Chinese translations if they are provided from China. Finally, if not satisfied, the requester has the option of rejecting the work of individual workers, at no cost.

Educating the turker with explicit and easy-to-understand instructions about the completion of the task is the first step towards quality. A sample translation HIT is shown in Figure 2. We provide detailed instructions on the HIT for both completion of the task and its evaluation. In our experience, explicit instructions about rejection criteria has shown to discourage turkers from providing sub-standard translations. We also set the workers qualification threshold to 85%, which guarantees only those workers who have had a success rate of 85% or above in the past hits. This has helped guarantee high quality translations and with validation and verification for cleaning up occasional human data entry errors reasonable quality data can be obtained.

4.2. Pricing

The interesting opportunity at crowd sourcing places like Mechanical Turk is that not only do we have access to human resources in abundance, but at low cost. There are workers willing to help at prices under 1 cent per translation. When working with a language pair like Spanish-English, pricing is not an issue due to the availability of Spanish speakers, but we imagine pricing to play a major role as we start exploring other language pairs where not many speakers of the language can be found on the web. Amazon also provides a feature where an additional bonus can be given to a worker for completion of a HIT to satisfaction. Such discretionary amounts result in building an expert base for the task.

4.3. Countermeasures for Gamers

With availability of online translation systems like Google translate, Yahoo translate (Babblefish), translation tasks in crowd-sourcing become easy targets for cheating. This defeats the purpose of the task, as our corpus would then be biased towards some existing automatic MT system. Its cheating as it is done in spite of explicit instructions not to do the same. It is extremely important to keep gamers in check, as not only do they pollute the quality of the crowd data, but their completion of a HIT means that it becomes unavailable to genuine turkers willing to provide valuable translations. We therefore collect translations from automatic MT services a priory and use these to match and block submissions from such users. As most free online MT services available for Spanish are of reasonably high quality, it is difficult to distinguish users that agree with these systems by chance versus users that game the task. We only reject users that have a significant overlap with automatic MT output and do not agree with gold standard data. For our experiments, we work with Spanish-English data where we also had gold standard translations for the input sentences. We use it to eliminate such gamers but languages without gold standard pose an interesting challenge.

5. Sentence Selection via Active Learning

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

5.1. Our Setup for Active Learning in MT

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 U_0 according to our selection strategy and retrieve the best scoring sentence or 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 every single sentence is computationally infeasible and may not have a significant affect 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)
        Request Human Translations for s_i
        S_k = $S_k \cup s_i$
    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}
end for

5.2. Sentence Selection

We have designed our sentence selection strategy to be independent of the underlying SMT system or the models. 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 phrase-based SMT systems. The basic units of

such SMT systems are phrases and therefore we measure informativeness of a sentence in terms of the phrases. Our strategy is to select sentences that have the most representative ngrams that have not yet been seen in the bilingual corpus. Representativeness or the 'density' $d(S)$ of a sentence is computed as a function of the unlabeled monolingual data as seen in Equation 5. We use $P(x/L)$ to represent the probability of a phrase x in the set of labeled sentences L. Similarly we use $count(x/L)$ to represent the counts of x in L. We also introduce a decay on the density of an ngram based on its frequency in the labeled data. The parameter λ is used to set the slope of this decay. This has shown to be particularly useful in our experiments. Novelty or 'uncertainty' $u(S)$ 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 ngrams of size upto 3. We refer to this strategy as density weighted diversity sampling strategy (DWDS).

$$d(S) = \frac{\sum_{x \in Phrases(S)} P(x/U) * e^{-\lambda \cdot count(x/L)}}{\|Phrases(S)\|} \quad (1)$$

$$u(S) = \frac{\sum_{x \in Phrases(S)} \alpha}{\|Phrases(S)\|} \quad (2)$$

$$\alpha = \begin{cases} 1 & x \notin Phrases(L) \\ 0 & \end{cases} \quad (3)$$

$$Score(S) = \frac{(1+\beta^2) d(S) * u(S)}{\beta^2 d(S) + u(S)} \quad (4)$$

6. Translation Selection

To ensure quality of translation output, each translation is requested from multiple turkers, in our case from three different translators. Translation Selection, therefore, is the task of selecting a best translation from among multiple translations received from the crowd.

6.1. Translation Reliability Estimation

We use inter-annotator agreement as a metric to compute translation reliability. The assumption here is that the more number of times a sentence is translated similarly by two or more translators, the more likely it is to be a correct translation. We notice a relatively good degree of agreement between turkers on a sample of 1000 sentences selected from a Spanish corpus that is translated by three different translators. About 21.1% of the time all three agree with each other, for 23.8% of the sentences two translators agree, and for the remaining 55.1% there was no agreement between translators.

We currently use exact matching to compare translations, which is not robust to variations in spelling or other language phenomena. We will be extending this to more flexible approaches like edit-distance or ngram overlap based methods for matching. In this regard, automatic MT evaluation metrics like BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) are promising and will be explored as future work. A more challenging task is to perform matching when there could be more than one perfectly valid translations for a given sentence.

6.2. Translator Reliability Estimation

The above approach of seeking multiple translations from turkers and using inter-annotator agreement works great in accounting for natural variability of translators and reducing occasional human error. However, this is expensive and may not be a viable long-term strategy. We would therefore like to identify reliable translators who are good at the given task of translation. This can help us vary our strategies and amortize the cost in future translations. Reliability of a translator is also useful in selecting a best fit translation for a sentence when there is no agreement between multiple turkers.

Given a worker w_k and a set of his/her translations $T_k = \{t_j^k\}$, we estimate reliability based on translations from other workers $T = \{t_j^n\}$ as shown in equation below.

$$rel(w_k) = \frac{\sum_{t_i^k \in T_k} \sum_{t_j^n \in T} \delta(t_i^k, t_j^n)}{\|T_k\|}$$

$$\delta(t_i^k, t_j^n) = \begin{cases} 1 & t_i^k \equiv t_j^n \\ 0 & \end{cases}$$

6.3. Selecting One Best Translation

We use both translation reliability and translator reliability to select the one best translation. We use a naive selection strategy that works well as seen in our results. We select the translation with highest translation reliability and solve ties by preferring translator with highest reliability. In future we will also be exploring other sophisticated methods for translator reliability estimation similar to (Donmez et al., 2009).

7. Experiments

We experiment our active learning approach and the effectiveness of crowd-sourcing techniques separately. We perform our experiments on the Spanish-English language pair for two reasons. Spanish is a popularly spoken language on the web and it is easy to find non-expert translators on the web using Mechanical Turk. Secondly, Spanish-English has been explored well in the machine translation community and we have baselines to compare with, as well as datasets that have been created by experts. This allows us to draw comparative performance analysis between crowd and expert quality.

We use the 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 were tuned using standard MERT (Och, 2003). Our development set consists of 343 sentences and the test set consists of 500 sentences.

7.1. Sentence Selection via Active Learning

7.1.1. Setup

We first test the performance of our active learning sentence selection strategy. We start with an initial system trained on

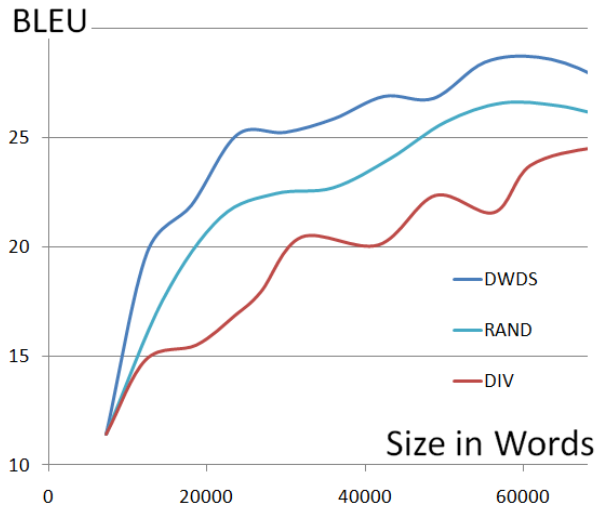

Figure 3: Spanish-English Sentence Selection results in a simulated AL Setup

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 the source side of the entire corpus. We simulate the human translation step in our experiment, as we already have access to the translations from the BTEC corpus. We then re-train, re-tune and test the system to complete the iteration.

7.1.2. Results

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 number of new ngrams presented in the sentence. We refer to this as a diversity strategy (DIV). This approach explicitly favors coverage, which is a desirable quality in machine translation. 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 - given an MT system, active learning strategy requires a smaller number of sentences to reach a desired performance thereby reducing cost of acquiring data.

7.2. Crowd-sourcing and Translation Selection

7.2.1. Setup

Unlike the simulated active learning setup discussed above, we performed this experiment for only two iterations. Each iteration of active learning experiment is similar to above, but instead of using already existing translations from the BTEC corpus, we crowd-sourced the data for translation via Mechanical Turk. Each sentence is presented to three different turkers for translation. Surprisingly, all the sentences were translated in less than 20 hours. In the first iteration 71 unique turkers participated to provide 3000 translations. A total of 17 man hours were spent among these turkers. The total cost for obtaining these translations was 45 USD. In the second iteration 64 turkers participated to provide 3000 translations. A total of 20 man hours were spent at a cost of 45 USD.

	Iterations	
System	**1**	**2**
crowd pick-rand	16.43	17.59
crowd translation-agreement	18.92	20.54
crowd translator-agreement	19.20	20.78
expert translations	19.21	21.29
crowd all-three	19.62	21.67

Table 1: Spanish-English Automatic MT Evaluation in Crowd-sourcing setup

7.2.2. Results

To evaluate effectiveness of crowd-sourcing for MT, we also conduct end to end MT experiments with data collected from the crowd. In Table 1 we show results from 'translation selection' as measured by BLEU (Papineni et al., 2002). Our experiments show the varying quality among online translators and hence the importance of translating the same sentence by multiple turkers. As discussed in section 6., when we use just translation reliability we already perform better than a random baseline where translations were picked as they came in from Mechanical Turk. Using 'translator' reliability along with 'translation' reliability performance improves further.

We also trained an MT system using translations from all three translators. This has proven to work quite well compared to selecting only one translation. This could be due to two reasons. Firstly, sub-sentential agreement across multiple translations reinforces the useful parts of a translation and suppresses noise. Secondly, multiple translations could contain paraphrases which can be useful (Callison-Burch et al., 2006). However obtaining multiple translations may prove expensive in the long run and hence is discouraged. As part of future work, we are exploring ways to use the worker reliability estimates in innovative ways for both designing the task and devising pricing strategies that further reduce overall cost of translation acquisition.

8. Conclusion

In this paper we proposed Active Crowd Translation (ACT), a new paradigm where active learning and crowd-sourcing come together to enable automatic translation for low-resource language pairs. Active learning aims at reducing cost of label acquisition by prioritizing the most informative data for annotation, while crowd-sourcing reduces cost by using the power of the crowds to make do for the lack of expensive language experts. We experimented and compared our active learning strategies with strong baselines and show significant improvements in translation quality even with less data. We used crowd-sourcing techniques for data acquisition at reduced costs using Amazon Mechanical Turk. Our experiments have shown that it is possible

to create parallel corpora using non-experts and with sufficient quality assurance, a translation system that is trained using crowd data approaches a system trained using expert data.

Acknowledgments

This research was partially supported by DARPA under grant NC 10-1326. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the DARPA. The first author would also like to thank Pooja Reddivari for help with creating graphic images and reviewing the paper.

9. References

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 17–24, Morristown, NJ, USA. Association for Computational Linguistics.

Chris Callison-burch. 2003. Active learning for statistical machine translation. In *Ph.D Proposal Edinburgh University*.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Singapore, August. Association for Computational Linguistics.

Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *KDD*, pages 259–268.

Matthias Eck, 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.

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

Gholamreza Haffari, 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.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: a study of annotation selection criteria. In *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Morristown, NJ, USA. Association for Computational Linguistics.

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

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of the HLT/NAACL*, Edomonton, Canada.

Philipp Koehn, 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*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *StatMT '07: Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Morristown, NJ, USA. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the EMNLP 2008*, pages 254–263, Honolulu, Hawaii, October.

Toshiyuki Takezawa, 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*.

Min Tang, Xiaoqiang Luo, and Salim Roukos. 2001. Active learning for statistical natural language parsing. In *ACL '02*, pages 120–127, Morristown, NJ, USA.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of EMNLP*.