Active Learning for Machine Translation in Low Resource Scenarios

Vamshi Ambati

CMU-SCS-10-xxx
November 9, 2010

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Jaime Carbonell (Chair)
Stephan Vogel (Co-chair)
Raj Reddy
Aniket Kittur
Kevin Knight, University of Southern California, ISI

Thesis Proposal

Copyright © 2010 Vamshi Ambati
Abstract

Corpus based approaches to automatic translation such as Example Based and Statistical Machine Translation systems use large amounts of parallel data created by humans to train mathematical models for automatic language translation. Large scale parallel data generation for new language pairs requires intensive human effort and availability of fluent bilinguals or expert translators. Therefore it becomes immensely difficult and expensive to provide state-of-the-art Machine Translation (MT) systems for rare languages.

In this thesis, we propose active learning to reduce costs and make best use of human resources for generating parallel data for MT systems. Active learning approaches help us identify sentences, which if translated have the potential to provide maximal improvement to an existing system. We then expand active learning to other MT relevant tasks such as word alignment, classifying monolingual text by topic, and a joint elicitation of two or more annotations for building domain specific MT systems.

As part of this thesis, we develop a new framework - Active Crowd Translation (ACT), a cost sensitive active learning setup for building MT systems for low-resource language pairs. Our framework will provide a suitable platform for involving disparately spread out human translators around the world, in a timely and sparingly fashion for rapid building of translation systems. We first explore the ACT paradigm with expert translators and then generalize to full-scale crowdsourcing with non-expert bilingual speakers. In case of Machine Translation, although crowdsourcing services like Amazon’s Mechanical Turk have opened doors to tap human potential, they do not guarantee translation expertise nor extended availability of translators. We address several challenges in eliciting quality translations from an unvetted crowd of bilingual speakers.
# Contents

Abstract 3

1 Introduction 9
   1.1 Statistical Machine Translation ................................. 10
   1.2 Active Learning .................................................... 12
   1.3 Dimensions of Active Learning ................................. 13
      1.3.1 Query Selection Frameworks ................................. 13
      1.3.2 Annotation Variety ........................................... 14
      1.3.3 Annotator ...................................................... 14
      1.3.4 Annotation Granularity ...................................... 14
      1.3.5 Operational Ranges ........................................... 15
   1.4 Thesis Proposal .................................................... 15
      1.4.1 Statement ...................................................... 15
      1.4.2 Hypotheses ...................................................... 15
      1.4.3 Proposed Work ............................................... 16
      1.4.4 Expected Contributions ..................................... 17
      1.4.5 Organization ................................................... 18

2 Literature Survey 19
   2.1 Statistical Machine Translation ................................. 19
   2.2 Active Learning .................................................... 20
      2.2.1 Cost-sensitive Active Learning .............................. 21
3 Active Learning for Machine Translation

3.1 Introduction

3.2 Active Sentence Selection for Parallel Data Creation
  3.2.1 Setup
  3.2.2 Evaluation
  3.2.3 Batch Mode Active Learning

3.3 Query Selection Strategies
  3.3.1 Data-Driven Selection
  3.3.2 Model-Driven Selection
  3.3.3 Decoding-Driven Selection

3.4 Word Alignment: A Second MT Task
  3.4.1 Active Learning Setup
  3.4.2 Semi-Supervised Word Alignment
  3.4.3 Query Selection Strategies

3.5 Experiments and Results
  3.5.1 Sentence Selection
  3.5.2 Word Alignment

3.6 Proposed Work
  3.6.1 Ensemble techniques for sentence selection

4 Multi Annotation Active Learning

4.1 Introduction

4.2 Focussed Domain Machine Translation
  4.2.1 Task 1: Sentence Classification
  4.2.2 Task 2: Sentence Translation
  4.2.3 Query strategy for multiple annotations
  4.2.4 Experiments

4.3 Proposed Work: Focused Domain MT
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Selection strategies for sentence classifier</td>
<td>48</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Sentence vs. Document Classification</td>
<td>48</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Cost Models</td>
<td>49</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Switching criteria</td>
<td>49</td>
</tr>
<tr>
<td>4.4</td>
<td>Proposed Work: Multi-Domain Machine Translation</td>
<td>50</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Problem setup</td>
<td>50</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Query strategies</td>
<td>50</td>
</tr>
<tr>
<td>4.4.3</td>
<td>Evaluation</td>
<td>52</td>
</tr>
<tr>
<td>5</td>
<td>CrowdSourcing for Machine Translation</td>
<td>53</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>54</td>
</tr>
<tr>
<td>5.2</td>
<td>Data Collection for MT</td>
<td>55</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Language Landscape of MTurk</td>
<td>55</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Challenges for Crowd-Sourcing and Machine Translation</td>
<td>56</td>
</tr>
<tr>
<td>5.3</td>
<td>Quality in Crowd</td>
<td>57</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Selective Approaches</td>
<td>58</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Collective Approaches</td>
<td>59</td>
</tr>
<tr>
<td>5.4</td>
<td>Cost Effective Elicitation</td>
<td>60</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Exploration vs. Exploitation</td>
<td>61</td>
</tr>
<tr>
<td>5.5</td>
<td>Experiments</td>
<td>62</td>
</tr>
<tr>
<td>5.5.1</td>
<td>Quality</td>
<td>62</td>
</tr>
<tr>
<td>5.5.2</td>
<td>Cost</td>
<td>63</td>
</tr>
<tr>
<td>5.5.3</td>
<td>Task based Evaluation</td>
<td>63</td>
</tr>
<tr>
<td>5.6</td>
<td>Proposed Work</td>
<td>64</td>
</tr>
<tr>
<td>5.6.1</td>
<td>New setups for Collaborative Translation</td>
<td>65</td>
</tr>
<tr>
<td>6</td>
<td>Summary and Timeline</td>
<td>67</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary of Work Done</td>
<td>67</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Active Learning for MT</td>
<td>67</td>
</tr>
<tr>
<td>6.1.2</td>
<td>Active Transfer Learning</td>
<td>67</td>
</tr>
<tr>
<td>6.1.3</td>
<td>Crowdsourcing for MT</td>
<td>68</td>
</tr>
</tbody>
</table>
6.2 Summary of Proposed Work ........................................... 68
   6.2.1 Active Learning for MT ........................................... 68
   6.2.2 Active Transfer Learning ........................................... 68
   6.2.3 Crowdsourcing for MT ........................................... 68
6.3 Schedule ................................................................. 69

Bibliography ............................................................... 71
Chapter 1

Introduction

We live in a world driven by information, and a generation that is making a push towards globalization. While information technology has made abundant progress to bridge the social, political and economic gap across culturally diversified population, a true sense of communication and information exchange is still impeded by the language barrier. The current world can be seen as a batch of communicating islands, at best. We want people to maintain their cultures, preserve their languages and so for the foreseeable future data will be produced in local languages. We want intelligent systems that interpret this content and deliver information where needed.

Machine translation has remained an interesting problem of natural language processing, aiming to bridge this exact language barrier, by building machines that can automatically translate between human languages. The state-of-the-art approaches to Machine Translation (MT) are data-driven requiring voluminous translation corpora. Given a large set of sentences in a source language and the equivalent translations in a target language, referred to as parallel corpora, the underlying algorithms learn word level dictionaries and other phrase patterns to support the translation of an unseen sentence.

Typically it has been the case that, with provision of more such translation data, the systems improve in quality. But, large scale parallel data generation is an onerous task requiring intensive human effort and availability of bilingual speakers. Consequently, much of the efforts made and progress seen has been confined to majority languages, where 'majority' can be interpreted as languages with large number of speakers or funding or political interest. However, a lot of languages in the world do not enjoy this status, which I refer to in this thesis as 'low-resource' languages.

Unlike majority language translation, where systems are more general, low-resource language translation systems, when built, are driven by an immediate need of a small group of audience and are therefore focused around a problem. For instance, projects that build
MT systems for African languages in order to aid the rehabilitation of refugees in the United States, or communication bridging translation tools for disaster relief volunteers. Such projects, which are of importance for humanitarian causes are also very time critical and typically have pre-specified, limited budgets. One cannot afford the translation of a million sentences to train high accuracy systems, neither can we wait for the time taken for data entry. Hence the need of the day is to build algorithms that provide usable translation systems at a very low budget and take less time for development.

In this thesis we resort to Active Learning (AL) techniques for building MT systems for minority languages. In a complex system like MT, different models combine forces to produce the final translation and therefore the annotations are multiple, structural and could be of different types. Depending upon the MT system and the paradigm, the resource requirements for translation may also encompass annotations such as morphological analysis, named-entity tagging, part-of-speech tagging, sense disambiguation, syntactic parses, semantic analysis, etc. We will focus on translation data acquisition and word-alignment, but our techniques should be useful and generalize to other types of annotations as well.

Traditionally, active learning strategies have been applied in learning a classifier, where the annotation is of a single data-type - the class label. We extend the traditional setup of active learning which elicits a single kind of annotation to improve the quality of a particular task to multiple annotations for different tasks.

Yet another factor in creating parallel data for MT is the non-availability of speakers of the language-pair. We use crowdsourcing techniques via social market places like Amazon's Mechanical Turk\(^1\) or web-scale labeling games in the spirit of ESP [Ahn and Dabbish, 2004] or ReCaptcha [Ahn et al., 2008] to collect data. However, crowd sources are inherently unreliable, so we will investigate proactive methods for estimating reliability and for combining results via weighted voting, jointly optimizing sample-selections, annotation-type selections, source-selection and multi-source combination strategies.

1.1 Statistical Machine Translation

Active Learning is dependent upon the definition of the methods and the model representation and so in this section we introduce the MT approach and the model we will optimize in the rest of the chapter. We discuss models pertaining to finite state methods, and later in the thesis we will extend them to context-free based approaches.

Machine Translation (MT) can be seen as a function, \( f : S \rightarrow T \) that transforms a sentence in a source language \( S \) to an equivalent translation into a target language \( T \). Learning such a function is a pattern recognition problem that involves modeling,

\(^1\)http://www.mturk.com/mturk/
1.1 Statistical Machine Translation

parametrization and parameter estimation. The task of translation can be seen as producing the most likely target sentence $t$ for a source sentence $s$ under a particular model $P_{\theta}(s/t)$, where $\theta$ can be any parameterization of the model.

$$s^* = \arg\max_s P_{\theta}(s/t)$$

For the most part of the thesis, we will work with phrase-based statistical machine translation (PB-SMT) system. The parameterization of such a system is motivated and focused in the generalization power expected of the model without sacrificing quality. A good unit of generalization that encapsulates local context, important for cohesive translation, are phrases. Phrase-based approaches rely on the existence of techniques for alignment at word-level, which are typically performed by the IBM models [Brown et al., 1993]. A large corpus with reasonably accurate alignment is a good starting point for applying extraction techniques for producing a translation equivalence table of phrase mappings, forming our model parameters. Parameter estimation of these parameters is a problem that has been at best addressed as relative likelihoods as provided by the data. The IBM models also provide other rich parameters in the form of lexical translation tables, and distortion parameters that are dependent upon the model assumptions and are an important by-product of the translation equivalence computation. These, in conjunction with the phrase table form the submodels of translation which are often linearly combined as shown below, the weights of which are optimized for translation performance on a held-out dataset.

$$P(e/f) = \frac{1}{Z(\lambda)} \prod_{i=1}^{m} h_i(e, f)^{\lambda_i}$$

The space as defined by the model $P_{\theta}(s'/t)$ of a PB-SMT consists of a number of submodels or feature functions $h_i(e, f)$ computed either from the source $s$, or target $f$ or both $(e, f)$. Some of the features observed in phrase based systems are as follows:

- $p_w(s_i/t_j)$: Lexical translation probability of the source word $s_i$ being a translation equivalent, given a target word $t_j$
- $p_w(t_i/s_j)$: Lexical translation probability of the target given source.
- $phr(s_i/t_j)$: Phrase translation probability of the source phrase $s_i$ being a translation equivalent, given a target word $t_j$
- $phr(t_i/s_j)$: Phrase translation probability of the target given source.
- $lm(t)$: Language model score for the target segment $t$. 
1.2 Active Learning

Supervised learning has become widely used and successful paradigm in natural language processing tasks. Most approaches rely on large scale labeled data availability for better accuracy rates. We will address the problem of selecting the most informative examples from unlabeled data in order to reduce human effort via ‘Active Learning’ methods. In active learning, the learner has access to a large pool of unlabeled data and sometimes a small portion of seed labeled data. The objective of the active learner is then to select the most informative instances from the unlabeled data and seek annotations from an oracle, which it then uses to retrain the underlying supervised model for improving performance. This continues in an iterative fashion for convergence - which typically is a threshold on the achievable performance before exhausting all the unlabeled data set. Since we consider the most informative unlabeled instances for annotation at every iteration, we always spend to improve the efficiency of the task.

Machine Translation (MT) requires labeled data in the form of translated sentences pairs where a translation can be seen as an annotation of a monolingual sentence. Similarly, other annotations tasks exist for MT such as word alignment, topic identification etc. Figure 1.1 shows a standard active learning setup for creating parallel corpus for an MT system, which can largely be generalized for other annotation tasks as well.

![Figure 1.1: Active Learning setup](image)

The main components of any active learning setup are therefore, the unlabeled data, a selection criteria and an oracle to annotate the selected data. We always assume availability of a large pool of unbiased collection of unlabeled data to select from. This is typically called a pool-based active learning. Finally an important aspect that we will not address in this thesis is the stopping criteria for active learning process. We assume that the process is
continued until a desired accuracy level is accomplished. For a more detailed description of active learning and its applications, we refer the reader to [Settles, 2009].

1.3 Dimensions of Active Learning

In this section we try to understand some of the dimensions of active learning that make it an interesting and challenging problem. We are not aware of work highlighting these dimensions in active learning. We will study the following dimension jointly and in isolation. Throughout the thesis, we will instantiate these in the context of MT and discuss them in further detail.

1.3.1 Query Selection Frameworks

Query strategies weight the importance of the unlabeled instances with respect to a given model and prioritize them for annotation. As we will study in the rest of this thesis, different query strategies have varied characteristics that can be exploited for a particular task and data set. So the choice of the algorithm is crucial to the benefits reaped. The following are two of some of the query frameworks which will be explored in this thesis.

Informativeness

Strategies in this primarily compute the representativeness of the instance among its pool. Selecting such instances provides better coverage over unseen future instances.

\[
\text{score}(s) = \sum_{s=1}^{\lvert U \rvert} \text{sim}(s, s') \times \frac{1}{\lvert U \rvert}
\]  

(1.1)

Uncertainty

Ability of scoring the instances with an uncertainty function reveals the weaknesses of the model. Sampling instances to fill in for this deficiency is expected to improve the model. In most structural learning problems the model takes the form of a probability distribution over the possible targets conditioned upon the input.

\[
\text{score}(s) = \theta(t/s)
\]  

(1.2)
1.3.2 Annotation Variety

Active learning approaches optimize the selection of samples for a particular labeling task involving a single kind of annotation. The choice of annotation constraints the data we select from, the methods for scoring instances, the expertise area of the annotator. For machine translation there is a wide variety of possible tasks that can benefit the translation system, some of which include - parallel data creation, alignment of word links in a sentence pair, rich tagging of words, syntactic tree annotation etc. The choice of annotation also influences the improvement in quality of the supervised algorithm. An interesting aspect to explore is the combination of multiple annotations and the joint select of the annotation variety and the instance.

1.3.3 Annotator

At the core of the learner is an external oracle that can provide answers/annotations to the selected dataset. One of the assumptions that has lately been questioned is the availability of such an expert and the non-fallible, infatiguable nature of the annotator. For most tasks, it is unlikely that such an oracle might exist, and if it does is so expensive that the cost of annotation becomes quickly prohibitive. Now, assume the highly expensive expert is replaced with a very cheap non-expert who is correct in only 70% of the cases. Can the learning algorithms deal with such noise? A crowd of non-experts may sometimes still be cheaper and more accessible than an expert. Active learning algorithms need to be aware of the budget and accuracy of the annotator in order to decide when and who and how many annotators to consult for annotating an instance.

1.3.4 Annotation Granularity

Another dimension that is often ignored in acquiring labels for classification tasks is the granularity. Labels in classification are often binary and thus can not further be decomposed. In structural learning though, the annotation granularity of the label is legitimate. For example, a partial sentence translation is still useful when compared to no translation. Seeking a phrase translation may be an easier task, that can be completed by more annotators, and has value in the training of an MT system. A cost-sensitive active learning algorithm can choose to fall back to fine-granular annotations to save on budget or utilize its human resources judiciously.
1.3.5 Operational Ranges

A dimension that has been ignored or not fully explored in most literature is that of an 'operating range'. Availability of initial labeled data (often termed as 'seed') changes the choice of the other dimensions of an active learning algorithm. For instance, when dealing with absolutely no labeled data, one might argue that a random selection of data may work well in quickly establishing a decision boundary. A similar argument can be applied to the choice of query sampling strategy when a lot of labeled data is available and only limited unlabeled data is streamed. We term these as 'operating ranges', the exact definition of which is dependent on the application. We will refer to these as 'context-aware query strategies' for active learning.

1.4 Thesis Proposal

Building Machine Translation systems for low-resource languages has a noble intention, useful for many. Corpus based approaches to automatic translation like Example Based and Statistical Machine Translation systems are the state-of-the-art. However, these approaches require significantly large amounts of resources, primarily in the form of parallel data to train mathematical models. Active learning addresses the situation when there is a paucity of labeled data such as bilingual parallel text, but unlabeled data such as monolingual untranslated text is available in abundance, and obtaining labels (e.g. topics, translations, alignments) requires extensive and expensive human effort. For most language pairs, professional translators are in fact expensive and too scarce and so building MT systems becomes a distant dream. We believe that the contributions from this thesis will provide better understanding into faster, cheaper and effective ways to building MT systems for such low-resource languages.

1.4.1 Statement

This thesis explores active learning for building statistical machine translation systems for low-resource languages and in the process, extends the fundamental framework of active learning in several dimensions which are explored within the context of several language pairs.

1.4.2 Hypotheses

- Active learning will reduce human costs in producing resources for corpus-based translation systems, with significant savings in the case of low resource language pairs.
• Combination of active learning tasks (e.g. topicality and translation) can further improve end-to-end translation efficiency, cost effectively

• Crowdsourcing and associated algorithms will help us reach large number of non-expert bilinguals and further reduce the cost of building MT systems

1.4.3 Proposed Work

The complex nature of machine translation task, poses several challenges to the application of active learning, more so in the context of low-resource languages. We performed experiments in various aspects of active learning and crowdsourcing applied to MT and study more in the due course.

We started this thesis with an exploration of active learning as a right fit for building low-resource machine translation systems. We identified a number of dimensions that are left unexplored in active learning, and argue that MT is an appropriate test-bed to experiment and validate some of these dimensions both independently and jointly. From the view of an MT system, we will also categorically study the different query strategies based upon the resource they use in selection. While some strategies only make use of the unlabeled data and parallel data for selection, making them MT paradigm agnostic, others use the underlying model representation or the hypothesized output, making them more closely tied to the model representation and limiting them to particular MT approaches, ex: SMT or EBMT etc.

We explored a variety of query strategy frameworks for selecting the optimal set of sentences to be translated by humans to aid training of an MT system. Some of the techniques we have looked at are data-driven and focus on the density and diversity statistics for selection. We will then explore novel ways of combining these strategies as static and dynamic ensembles. These experiments will be carried out for multiple language pairs to test both feasibility and portability. We also applied active learning to a second task in MT - word alignment, and have shown improved performance at less human effort.

We then propose to explore the concept of ‘operating ranges’ for active learning. Similar to research in [Donmez et al., 2007], we also observe that in most related work the application of active learning strategies is done in an unified fashion all along the learning curve. However, as the data acquired in active learning process increases, the MT system moves from being a low-resource to a medium-resource and then eventually becomes a high-resource language. Query strategies need to be introspective of the resources available to the system. We will observe for the existence of these operating ranges in a learning curve and study the extension of introspective query strategies for optimal data selection under varied data scenarios.

A second direction that this thesis will explore is that of extending active learning to
handle multiple annotations. Active learning traditionally selects instances for a single kind of annotation task. In Machine translation several different kinds of data and annotation can be provided for the improvement of the underlying model, each associated with respective cost models. We will address this in the context of building multi-domain translation systems, where domain information and translation are two different annotations and we will explore the benefits from combining them under a single learning algorithm. The learning strategies need to be aware of the annotation variety and optimally select the annotation and the instance together to trade-off between cost and benefit.

A third direction that we will explore in greater depth is the assumption of expert oracle availability for providing annotations in an active learning setup. We argue that this is not only expensive but also less feasible for low-resource languages where finding speakers of the languages is a difficult proposition. We propose crowdsourcing as an alternative data collection. The challenges with using this paradigm is the affect of quality of the data, inspired by the techniques in proactive learning we will both implement existing and propose new algorithms for dealing with non-experts in MT. We will also study empirically the feasibility of using such data in training real world MT systems.

Finally, we will unify these ideas under a cost-sensitive framework which we call ‘active crowd translation’, geared towards building low-resource language translation systems.

1.4.4 Expected Contributions

The major contribution of this thesis is the development and evaluation of a new framework for building machine translation systems for low-resource languages. We call this Active Crowd Translation (ACT). Along the path, we expect to make the following contributions:

- Development of active learning algorithms for multiple tasks in machine translation of low-resource languages like - parallel data creation, word alignment.

- Extension of traditional ‘single-annotation’ driven active learning to handle multiple annotations in the context of multi-domain machine translation.

- Application of crowdsourcing techniques for building parallel corpora and other resources required for translation, with a focus on controlling for quality of the data.

- A budgeted cost motivated framework for Active Crowd Translation, with a deeper understanding of approaches to building MT systems for different language pairs in diverse data scenarios.
1.4.5 Organization

- Chapter 2 provides a literature survey of the three main broad areas supporting this thesis - active learning, crowdsourcing and machine translation.

- Chapter 3 discusses application of active learning to two tasks in Machine Translation - sentence selection and word alignment. We discuss several query selection strategies for both the tasks.

- Chapter 4 provides a novel extension to the traditional single-annotation focussed active learning to a multi annotation setup. We discuss the application of multiple annotation active learning in the context of building domain specific translation systems.

- Chapter 5 introduces crowdsourcing as a viable option for eliciting parallel data for building MT systems. We analyze several issues in this new area of research and provide algorithms for making use of the selective and/or collective wisdom of the crowds.

- Chapter 6 discusses some conclusions and a timeline for the rest of the thesis work.
Chapter 2

Literature Survey

In this chapter we will provide some background and survey literature pertaining to the three major fields that are relevant to this thesis - Statistical Machine Translation, Active Learning and Crowdsourcing.

2.1 Statistical Machine Translation

In recent years, corpus based approaches to machine translation have become predominant, with Phrase Based Statistical Machine Translation (PB-SMT) [Koehn et al., 2003] being the most actively progressing area. While PB-SMT improves traditional word based machine translation approaches by incorporating more contextual information in the form of phrase pairs, it still has limitations in global block level reordering of phrasal units. Such reorderings can be captured by knowledge about the structure of the language. Recent research in syntax based machine translation [Yamada and Knight, 2001] [Marcu et al., 2006] [Chiang, 2005] incorporates syntactic information to ameliorate the reordering problem of phrasal units. Some of the approaches operate within the resources of PB-SMT and induce hierarchical grammars from existing non-syntactic phrasal units, to provide better generality and structure for reordering [Chiang, 2005] [Wu, 1997]. Other approaches use syntactic analysis of sentences on one side of the corpus to induce grammar rules [Galley et al., 2004] [Yamada and Knight, 2001] [Venugopal et al., 2007].

Most approaches that incorporate linguistic syntax start with word level alignments and a parse tree for one side of the language pair, and obtain phrase tables and hierarchical translation rules driven by the syntax. We call this the ‘TnS’ setting, where we have the tree on one side and only string on the other. While this has indeed proven successful [Yamada and Knight, 2001] [Marcu et al., 2006], it has been shown that the word alignments which are usually extracted using syntactically uninformed generative models are not optimal for
the syntactic phrase extraction problem [DeNeefe et al., 2007, DeNero and Klein, 2007]. Some approaches [Crego and Habash, 2008, Fossum et al., 2008] have been proposed to modify the word alignments in ways that make them more amenable to building syntactic models.

The hierarchical grammar rules can be understood as context-free re-writing systems with left-hand side (LHS) and a right-hand side (RHS). The LHS of a particular rule can correspond to multiple RHS, resulting in an ambiguity factor that is typically resolved during decoding as a search problem. SCFG independence assumptions reflect upon the ambiguity factor seen in the grammars. Syntax based translation systems usually resort to a number of syntactic features and scoring techniques to resolve this ambiguity. However, SCFG translation models that are often learnt from parallel corpus and syntactic parse trees under certain independence assumptions are often so generalized that it becomes quite difficult to explore a meaningfully small space to obtain the right translation hypothesis.

2.2 Active Learning

Over the past two decades, supervised learning in structured spaces has been quite successful in syntactic analysis problems in natural language processing. With the initial success of such techniques in statistical parsing [Charniak, 2000], sequence labeling tasks [Brill, 1992], more researchers have been focusing on learning from data to solve NLP tasks. These learning techniques exploit large amounts of annotated data to learn models that can perform linguistic analysis on unseen data. The quantities of the annotated data are far from being sufficient for the majority of languages. Acquiring such supervised linguistic annotations for a language is important for natural language processing and it usually involves significant human efforts. Researchers have quickly realized this problem and are moving to ways of reducing the burden of annotation, so as to reduce the time to production of their tools. Active Learning is a field of Machine Learning that studies selective sampling of most informative examples for annotation. In this section, we survey the literature for active learning applications in Natural Language Processing (NLP). For a more general survey of Active Learning we refer to [Settles, 2009].

Active learning has been applied successfully in the area of Computer Vision for collecting data to improve the object identification task [Holub et al., 2008], face recognition [Hewitt and Belongie, 2006] among others. In speech technologies, data collection for building automatic speech recognition systems benefited by application of active learning techniques [Hakkani-tr et al., 2002]. Annotated data is used by most supervised and semi-supervised learners in NLP to solve two kinds of problems, classification tasks such as text classification [McCallum and Nigam, 1998], and structure prediction tasks such as statistical parsing [Charniak, 2000], sequence labeling [Brill, 1992].
2.2 Active Learning

[Thompson et al., 1999] applied uncertainty based active learning to the task of training a semantic parser. They consider two systems CHILL [Zelle and Mooney, 1996] and RAPIER that map sentences to their semantic representation. They use an uncertainty criteria which selects sentences that are not parsed by the systems. [Roth and Small, 2006] present a margin-based method for active learning in structured outputs. In particular they apply this to the task of semantic role labeling which is the task of identifying and labeling the semantic arguments of a predicate. In the case of margin-based classifiers, uncertainty translates to the distance from the hyper-plane. They explore the tradeoff between selecting instances based on a global margin or a combination of the margin of local classifiers.

Grammar induction is the task of inferring grammatical structure of a language. Research in statistical parser induction [Charniak, 2000] shows that given enough labeled data with good quality annotations, grammar induction can be done with reasonable accuracy. [Hwa, 2004] use selective sampling to minimize the amount of annotation needed for corpus based grammar induction. They use uncertainty based selective sampling techniques, where the uncertainty is computed as the length of the sentence or a tree entropy metric. They select sentences for annotation that have a uniform parse tree distribution. [Steedman et al., 2003] experiment uncertainty based metrics as provided by their parser to select sentences and show that improved accuracy can be achieved by annotating fewer sentences. [Baldridge and Osborne, 2003] applies similar approaches to selection of sentences for grammar induction under a HPSG grammar formalism.

2.2.1 Cost-sensitive Active Learning

We may have to build upon recent work in multi-task active learning that looks at selection of corpus for two relatively different tasks [Reichart et al., 2008b]. They experiment with different ways of combining the ranking of unlabeled data as optimized to each of the tasks. [Krause and Horvitz, 2008] applies cost sensitive learning to the task of asking questions for learning privacy preferences of an individual, while [Melville et al., 2005] look at application in feature annotation vs. instance annotation. In translation optimizing cost for low-resource languages is important and we will be working on optimizing cost in a multi-task scenario: translation vs. word-alignment.

2.2.2 Machine translation and active learning

For statistical MT, 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. Chris Callison-Burch provided a research proposal that lays out a plan of action for active learning for SMT [Callison-burch, 2003]. However it lacked any further experimentation and results. 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. [Haffari and Sarkar, 2009] discuss an active learning task of introducing a new language pair into an existing multilingual set of parallel texts with a high quality MT system for each pair. This novel setup is applicable when working with multi-language parallel corpus, and they exploit it to propose new sentence selection strategies and features.

[Eck et al., 2005] apply active learning as a weighting scheme to select more informative sentences in order to port their MT system to low resource devices like PDAs. The informativeness of a sentence is estimated using unseen n-grams in previously selected sentences. [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.

[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. They assume the existence of target-side translations along with the source-side sentences to compute features useful in estimating utility of a sentence pair. [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.

In an effort to provide translations for difficult-to-translate phrases, [Mohit and Hwa, 2007] propose a technique to classify phrases as difficult and then call for humans to translate them. This can be seen as an active learning strategy for selection at phrase level. [Kato and Barnard, 2007] implement active learning and semi-supervised approaches to build MT systems for language pairs with scarce resources. They show results with very limited data simulated with languages like Afrikaans, Setswana, etc.

### 2.3 Crowd Sourcing

With the advent of online market places such as Amazon Mechanical Turk, it is now easier to reach annotators 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-
2.3 Crowd Sourcing

sourcing’ for acquisition of supervised data [Snow et al., 2008] and conducting user studies [Kittur et al., 2008] where annotation tasks are farmed out to a large group of users on the web utilizing micro payments. [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.

In case of Machine Translation, although services like Mechanical Turk have opened doors to tap human potential, they do not guarantee translation expertise nor large-volume availability of translators. Here proactive learning [Donmez and Carbonell, 2008] adds other useful dimensions to active learning in a crowd-sourcing scenario - such as coping with information sources of variable cost and variable reliability of information sources [Donmez et al., 2009, 2010] and jointly optimizing the selection of instance (what to translate or align) with the selection of source (which individual or group to rely upon) modulated by cost (more reliable annotators may be more expensive). These techniques become increasingly beneficial as we move to a crowd-sourcing scenario for eliciting annotations. In our case, the trade-off may be between an expert translator and multiple cheaper translators of unknown reliability.
Chapter 3

Active Learning for Machine Translation

All data is not made equal and all sentences do not contribute equally to the performance of an MT system. In this chapter we discuss our approaches to actively prioritizing and selecting instances from a large pool of data so as to benefit the translation system. We will focus on two tasks that are important to a general MT system - parallel data creation, and word alignment. We show that both tasks can benefit from application of active learning.

3.1 Introduction

Active learning is appropriate for tasks where unlabeled data is readily available, but obtaining annotations is expensive. While there has been much work on active learning for classification tasks like document categorization [McCallum and Nigam, 1998] , active learning for structural learning tasks has received considerably less attention. However, due to the structured nature of labeling tasks, annotating these instances can be rather tedious and time-consuming, making active learning an attractive technique. Recently successful attempts have been made in applying active learning to sequence labeling problems like part-of-speech tagging, entity tagging and other tasks like parsing [Steedman et al., 2003], information extraction [Roth and Small, 2006] and etc.

In MT, only a few successful attempts have been made, but mostly using uncertainty sampling approaches [Haffari et al., 2009, Eck et al., 2005]. The methods of evaluation and applied strategies however make it difficult to draw conclusions on the comparative effectiveness of various approaches. In this section we make an attempt to establish a setup for active learning in MT, and categorize multiple query strategies based on the resources they use and effectiveness. We compare them under multiple data settings and evaluation
metrics to study their effectiveness on a varied set of language pairs. We also discuss batch selection mechanisms and their importance in the MT setup. We discuss our novel approach to active sentence selection that is a combination of the informativeness of the sentence, uncertainty and batch diversity.

3.2 Active Sentence Selection for Parallel Data Creation

3.2.1 Setup

We first discuss our general framework for active learning in SMT followed by the selection approaches. 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 shown in Algorithm 1.

\section*{Algorithm 1 Active Learning for SMT}

\begin{algorithmic}
  \STATE Given Labeled Data Set : \( L_0 \)
  \STATE Given Unlabeled Data Set: \( U_0 \)
  \FOR {\text{for } k = 0 \text{ to } T} 
    \FOR {\text{for } i = 0 \text{ to } N} 
      \STATE \( s_i = \text{Query}(U_i, L_i) \) 
      \STATE \( t_i = \text{Human Translation for } s_i \) 
      \STATE \( S_k = S_k \cup (s_i, t_i) \) 
    \ENDFOR 
    \STATE \( U_{k+1} = U_k - S_k \) 
    \STATE \( L_{k+1} = L_k \cup S_k \) 
    \STATE Re-train MT system on \( L_{k+1} \) 
  \ENDFOR 
\end{algorithmic}

3.2.2 Evaluation

The effectivieness of active learning approaches is typically measured in terms of the cost savings that result due to the prioritized selection of data. Most active learning approaches
consider a uniform cost of elicitation for the instances, which clearly is not true for MT. Estimating the cost model for translation is complicated for two reasons. Sentences, which are the instances here, are of different lengths to start with and a long sentence takes more time to translate than a short sentence. However, if we consider time taken to translate a sentence as a metric of cost, we are faced with the issue of accurate measurement of time in an online scenario. This situation is exacerbated by the fact that humans work at different paces. We therefore use word count as our cost model and try to optimize the number of words translated for improvement in translation quality. The cost of translation per sentence is treated as the sum of cost of translation of the consisting words. Although we acknowledge that cost of words may not be uniform, with some words more difficult to translate than other, in this work we assume a uniform cost of word translation across a particular domain, similar to industry quality human translation costs. To measure improvement in translation quality, we use automatic translation metrics like BLEU [Papineni et al., 2002] and METEOR [Lavie and Agarwal, 2007].

3.2.3 Batch Mode Active Learning

In most active learning research, queries are selected online, i.e., one at a time. In machine translation and other structure prediction tasks, the training time required to induce a model is slow or expensive making it difficult to work with labeling single units. While annotation was considered a cumbersome process, working at much slower speeds than model training, with the advent of web platforms like Mechanical Turk (MTurk) it is now possible to do large scale labeling in parallel using large crowds. Batch model learning therefore is an interesting alternative to online active learning.

The challenge in batch mode learning is however to select the batch appropriately so that the selected instances have minimal overlap among each other, in order not to introduce redundancy. Also, the selection of each instance influences the underlying model, and in turn effects the scores of the future instances to be selected. When the retraining of the model is not possible due to computational complexity, the selection of instances is sub-optimal due to the staleness of the model parameters. In our work introduce a decay parameter based on the number of times the instance has been seen in the batch previously. As will be discussed later, this component can be coupled with any of our query strategies inorder to combat the batch selection problem.

3.3 Query Selection Strategies

In this section we discuss our various query strategies for sentence selection. We categorize the multiple strategies into three categories based on the resources used in score
3.3.1 Data-Driven Selection

Data-driven sentence selection strategies only use the monolingual data \( U \) and the bilingual parallel corpus \( L \) to select sentences. Approaches in this category are independent of the underlying MT system and so are agnostic to the model representation. This makes our approach applicable to any corpus-based MT paradigm and system, even though we test on SMT in this thesis.

Density based approach

An important criteria in translation is to train a model in order to decode an unseen future sentence. Therefore, all else equal, we wish to select sentences that are most similar to the rest of the data, as they form a representative sample of the corpus. Moreover, sentences that are drawn from the distribution of available sentences are more likely to be able to provide coverage for unseen sentences.

\[
s = \arg \max_s \text{sim}(s, U)
\]  

(3.1)

The basic units of an SMT system are phrases and therefore we measure the similarity across sentences in terms of the consisting phrases. Our scoring strategy is shown in Equation 3.2. We select sentences that have the most representative n-grams of the bilingual corpus. Representativeness or the ‘density’ of a sentence is computed by \( P(x|U) \) as relative likelihood estimates of the n-gram 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. We use a decay parameter instead of completely ignoring already seen phrases for two reasons. Firstly, phrases that are dense typically are polysemic with a higher translation fan-out, which means they are translated into different target sides in different contexts. A gradual decay allows us to sample the phrase multiple times allowing us to capture the spectrum of variations. Secondly, the decay also works against redundancy within the batch allowing us to do a batch selection.

\[
dden(s) = \frac{\sum_{x \in \text{Phrases}(s)} P(x|U) \cdot e^{-\lambda \text{count}(x|L)}}{||\text{Phrases}(s)||}
\]

(3.2)

\[
P(x|U) = \frac{\text{count}(x)}{\sum_{x' \in \text{Phrases}(U)} \text{count}(x')}
\]

(3.3)
3.3 Query Selection Strategies

Diversity based approach

We also design a strategy that favors sentences providing higher coverage. Diversity or novelty, is computed as the number of new phrases that a sentence has to offer. This can be computed from the labeled data as shown in Equation 3.4.

\[
\text{div}(s) = \sum_{x} \frac{\alpha}{|\text{Phrases}(s)|} \alpha = \begin{cases} 
1 & x \notin \text{Phrases}(L) \\
0 & \text{otherwise}
\end{cases}
\]  \hspace{1cm} (3.4)

Density Diversity Ensemble

Ensemble techniques tend to work better as they provide different perspectives of the value function [Melville and Mooney, 2004, Freund et al., 1997]. We therefore devise a density and diversity ensemble that favors dense and novel phrases, by computing the final score of a sentence as the harmonic mean of both the two metrics with a tunable parameter ‘\(\beta\)’, 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.

\[
\text{ddendiv}(s) = \frac{(1 + \beta^2)d(s) * u(s)}{\beta^2 d(s) + u(s)}
\]  \hspace{1cm} (3.5)

KL Divergence

This strategy selects a sentence that highly contributes to the KL-divergence between the labeled data distribution and the unlabeled data distribution.

\[
\text{kldiv}(s) = P(s|U) \log \frac{P(s|U)}{P(s|L)}
\]  \hspace{1cm} (3.6)

\[
s^* = \arg \max_s \text{kldiv}(s)
\]  \hspace{1cm} (3.7)

3.3.2 Model-Driven Selection

Unlike the data-driven approaches above, in the model-driven approaches, we make use of the model trained using the labeled data. Therefore model based approaches are specific to the MT system and the underlying model representation, which in our case is phrase based statistical machine translation system (PB-SMT). This strategy can be called an uncertainty sampling technique [Lewis and Catlett, 1994], quite popular in active learning.

Given a sentence pair we can compute several measure of uncertainty associated with them, such as entropy or confidence as shown in Equation 3.9. Uncertainty can be computed
under a certain model and therefore this technique is dependent upon the MT system and
the underlying model. We use the translation models trained so far to select these sentence
pairs. The translation models in SMT typically consist of phrase tables with bidirectional
translation scores and bidirectional lexicon tables that can be used to compute source-give-
target and target-given-source probabilities at word level. We use \( t2s \) to denote computation
using ‘target-given-source’ models and \( s2t \) to denote computation using ‘source-given-target’
models. We select the sentences that have a maximum cumulative entropy under the phrase
translation model as computed below, where \( \text{Phr}(x) \) denotes all the phrases present in a
sentence \( x \), \( \text{Trans}(x) \) denotes all the phrase translations of \( x \) and \( p(t|s) \) is the conditional
translation probability distribution of phrases.

\[
\text{cond}_\text{entropy}(s) = \frac{\sum_x \sum_y \text{Trans}(x) - p(y|x) \log(p(y|x))}{|\text{Phrases}(s)|} \\
s^* = \arg \max_s \text{cond}_\text{entropy}(s)
\] (3.8)

We use the score to prioritize the sentences with the highest entropy, effectively prefer-
ing a sentence containing more number of high uncertain phrases. The intuition is that
the translation of such a sentence will lead to maximal reduction of entropy of the current
translation model.

### 3.3.3 Decoding-Driven Selection

One of the indicators of translation quality is the conditional distribution of the target
hypothesis over the source \( P(t|s) \). For most decoding approaches, the target conditional
distribution is a criteria for inference, and is also the best possible indicator we have for
overall translation quality. We normalize the score in order for it to be comparable across the
source sentences. We select the sentences with low score for the \( \text{decode}_\text{score}(s) \), indicative
of the weakness of the model in translation of the sentence.

\[
\text{decode}_\text{score}(s) = \frac{1}{|\text{len}(s)|} \arg \max_t P_0(t|s) \\
s^* = \arg \min_s \text{decode}_\text{score}(s)
\]

### 3.4 Word Alignment: A Second MT Task

The success of statistical approaches to MT can be attributed to the IBM models [Brown
et al., 1993] that characterize word-level alignments in parallel corpora. Parameters of
these alignment models are learnt in an unsupervised manner using the EM algorithm over
sentence-level aligned parallel corpora. While the ease of automatically aligning sentences at the word-level with tools like GIZA++ [Och and Ney, 2003] has enabled fast development of SMT systems for various language pairs, the quality of alignment is often quite low, especially for language pairs like Chinese-English, Arabic-English that diverge from the independence assumptions made by the generative models. Increased parallel data enables better estimation of the model parameters, but a large number of language pairs still lack such resources.

Two directions of research have been pursued for improving generative word alignment. The first is to relax or update the independence assumptions based on more information, usually syntactic, from the language pairs [Cherry and Lin, 2006, Fraser and Marcu, 2007b]. The second is to use extra annotation, typically word-level human alignment for some sentence pairs, in conjunction with the parallel data to learn alignment in a semi-supervised manner. Our research is in the direction of the latter, and aims to reduce the effort involved in hand-generation of word alignments by using active learning strategies for careful selection of word pairs to seek alignment.

3.4.1 Active Learning Setup

We discuss our active learning setup for word alignment in Algorithm 2. We start with an unlabeled dataset $U = \{(S_k, T_k)\}$, indexed by $k$, and a seed pool of partial alignment links $A_0 = \{a_{ij}^k, \forall s_i \in S_k, t_j \in T_k\}$. This is usually an empty set at iteration $t = 0$. We iterate for $T$ iterations. We take a pool-based active learning strategy, where we have access to all the automatically aligned links and we can score the links based on our active learning query strategy. The query strategy uses the automatically trained alignment model $M_t$ from the current iteration $t$ for scoring the links. Re-training and re-tuning an SMT system for each link at a time is computationally infeasible. We therefore perform batch learning by selecting a set of $N$ links scored high by our query strategy. We seek manual corrections for the selected links and add the alignment data to the current labeled data set. The word-level aligned labeled data is provided to our semi-supervised word alignment algorithm for training an alignment model $M_{t+1}$ over $U$.

We can iteratively perform the algorithm for a defined number of iterations $T$ or until a certain desired performance is reached, which is measured by alignment error rate (AER) [Fraser and Marcu, 2007a] in the case of word alignment. In a more typical scenario, since reducing human effort or cost of elicitation is the objective, we iterate until the available budget is exhausted.
Algorithm 2 AL for Word Alignment

1: Unlabeled Data Set: \( U = \{ (S_k, T_k) \} \)
2: Manual Alignment Set : \( A_0 = \{ a_{ij}^k, \forall s_i \in S_k, t_j \in T_k \} \)
3: Train Semi-supervised Word Alignment using \( (U, A_0) \rightarrow M_0 \)
4: \( N \): batch size
5: for \( t = 0 \) to \( T \) do
6: \( L_t = \text{LinkSelection}(U, A_t, M_t, N) \)
7: Request Human Alignment for \( L_t \)
8: \( A_{t+1} = A_t + L_t \)
9: Re-train Semi-Supervised Word Alignment on \( (U, A_{t+1}) \rightarrow M_{t+1} \)
10: end for

3.4.2 Semi-Supervised Word Alignment

We use an extended version of MGIZA++ [Gao and Vogel, 2008] to perform the constrained semi-supervised word alignment. Manual alignments are incorporated in the EM training phase of these models as constraints that restrict the summation over all possible alignment paths. Typically in the EM procedure for IBM models, the training procedure requires for each source sentence position, the summation over counts from all words in the target sentence. The manual alignments allow for one-to-many alignments and many-to-many alignments in both directions. For each position \( i \) in the source sentence, there can be more than one manually aligned target word. The restricted training will allow only those paths, which are consistent with the manual alignments. Therefore, the restriction of the alignment paths reduces to restricting the summation in EM.

3.4.3 Query Selection Strategies

We propose multiple query selection strategies for our active learning setup. The scoring criteria is designed to select alignment links across sentence pairs that are highly uncertain under current automatic translation models. These links are difficult to align correctly by automatic alignment and will cause incorrect phrase pairs to be extracted in the translation model, in turn hurting the translation quality of the SMT system. Manual correction of such links produces the maximal benefit to the model. We would ideally like to elicit the least number of manual corrections possible in order to reduce the cost of data acquisition. In this section we discuss our link selection strategies based on the standard active learning paradigm of ‘uncertainty sampling’[Lewis and Catlett, 1994]. We use the automatically trained translation model \( \theta_t \) for scoring each link for uncertainty, which consists of bidirectional translation lexicon tables computed from the bidirectional alignments.
3.4 Word Alignment: A Second MT Task

Uncertainty Sampling: Bidirectional Alignment Scores

The automatic Viterbi alignment produced by the alignment models is used to obtain translation lexicons. These lexicons capture the conditional distributions of source-given-target \( P(s|t) \) and target-given-source \( P(t|s) \) probabilities at the word level where \( s_i \in S \) and \( t_j \in T \). We define certainty of a link as the harmonic mean of the bidirectional probabilities. The selection strategy selects the least scoring links according to the formula below which corresponds to links with maximum uncertainty:

\[
Score(a_{ij}|s_I^t, t^T_j) = \frac{2 \cdot P(t_j|s_i) \cdot P(s_i|t_j)}{P(t_j|s_i) + P(s_i|t_j)}
\]  

Confidence Sampling: Posterior Alignment probabilities

Confidence estimation for MT output is an interesting area with meaningful initial exploration [Blatz et al., 2004, Ueffing and Ney, 2007]. Given a sentence pair \((s_I^t, t^T_j)\) and its word alignment, we compute two confidence metrics at alignment link level – based on the posterior link probability as seen in Equation 3.10. We select the alignment links that the initial word aligner is least confident according to our metric and seek manual correction of the links. We use \( t2s \) to denote computation using higher order (IBM4) target-given-source models and \( s2t \) to denote source-given-target models. Targeting some of the uncertain parts of word alignment has already been shown to improve translation quality in SMT [Huang, 2009]. We use confidence metrics as an active learning sampling strategy to obtain most informative links. We also experimented with other confidence metrics as discussed in [Ueffing and Ney, 2007], especially the IBM 1 model score metric, but it did not show significant improvement in this task.

\[
P_{t2s}(a_{ij}, t^T_j|s_I^t) = \frac{p_{t2s}(t_j|s_i, a_{ij} \in A)}{\sum_{i} p_{t2s}(t_j|s_i)}
\]  

\[
P_{s2t}(a_{ij}, s_I^t|t^T_j) = \frac{p_{s2t}(s_i|t_j, a_{ij} \in A)}{\sum_{i} p_{s2t}(s_i|t_j)}
\]  

\[
Conf(a_{ij}|S, T) = \frac{2 \cdot P_{t2s} \cdot P_{s2t}}{P_{t2s} + P_{s2t}}
\]

Query by Committee

The generative alignments produced differ based on the choice of direction of the language pair. We use \( A_{t2s} \) to denote alignment in the source to target direction and \( A_{s2t} \) to denote the target to source direction. We consider these alignments to be two experts that have two different views of the alignment process. We formulate our query strategy to select
links where the agreement differs across these two alignments. In general query by committee is a standard sampling strategy in active learning [Freund et al., 1997], where the committee consists of any number of experts, in this case alignments, with varying opinions. We formulate a query by committee sampling strategy for word alignment as shown in Equation 3.14. In order to break ties, we extend this approach to select the link with higher average frequency of occurrence of words involved in the link.

\[
Score(a_{ij}) = \alpha
\]

where \( \alpha = \begin{cases} 
2 & a_{ij} \in A_{s2t} \cap A_{t2s} \\
1 & a_{ij} \in A_{s2t} \cup A_{t2s} \\
0 & \text{otherwise}
\end{cases} \)  

### Margin Sampling

The strategy for confidence based sampling only considers information about the best scoring link from Eq 3.13. However we could benefit from information about the second best scoring link as well. Earlier work has shown success in multi-class classification problems using such a ‘margin based’ approach, where the difference between the probabilities assigned by the underlying model to the first best and second best labels is used as a sampling criteria [Scheffer et al., 2001]. We adapt such a margin-based approach to link-selection using the \( \text{Conf}_1 \) scoring function discussed in the earlier sub-section. Our margin technique is formulated below, where \( \hat{a}_{ij1} \) and \( \hat{a}_{ij2} \) are potential first best and second best scoring alignment links for a word at position \( i \) in the source sentence \( S \) with translation \( T \). The word with minimum margin value is chosen for human alignment. Intuitively such a word is a possible candidate for mis-alignment due to the inherent confusion in its target translation.

\[
\text{Margin}(i) = \text{Conf}(\hat{a}_{ij1}|S,T) - \text{Conf}(\hat{a}_{ij2}|S,T)
\]

### 3.5 Experiments and Results

#### 3.5.1 Sentence Selection

We perform our experiments on the Spanish-English language pair. 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 were tuned using standard MERT [Och, 2003].
Our development set consists of 343 sentences. We currently report results on this set and
will be testing on other data sets in final version. The test set used consists of 500 sentences.

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 step in our
experiment, as we already have access to the translations from the BTEC corpus. We then
retrain, re-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 capture 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.
This is natural as this would be the manner in which we would provide data to an expert for
translation. 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 [Papineni et al., 2002] on
a held out dataset. As seen in Figure 3.1, some of our active learning strategies perform
better than the random baseline. Our most significant improvements come from the density
and diversity ensemble (dden-div). The difference between the curves is most prominent
in the earlier iterations indicating the importance of a density based sampling strategy for
very low resource scenarios. 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 uses less number of sentences to reach a desired accuracy thereby
reducing cost of acquiring data. This shows that we can achieve similar performances by
spending lot less for translation.

We also tested end-to-end MT performance on two other language pairs. For the
Japanese-English language pair we use the BTEC travel corpus released under the IWSLT.
The parallel corpus consisted of 160k sentence pairs and the results are shown in Figure 3.2.
For Urdu-English, we use the NIST MT Eval dataset for 2008, with 50k sentence pairs. The
results from testing it on a standard newswire test set released under the dataset can be seen
in Fig 3.3. In both the language pairs, we notice quite some improvement with different
query strategies like kl-div, dden-div, and significant results for the dden-div strategy across
all language pairs.

3.5.2 Word Alignment

For word alignment we performed experiments to show that active learning can help select
the most informative alignment links that have high uncertainty according to a given
Figure 3.1: Spanish-English Sentence Selection results in a simulated AL Setup

Figure 3.2: Japanese-English Sentence Selection results in a simulated AL Setup
3.5 Experiments and Results

Figure 3.3: Urdu-English Sentence Selection results in a simulated AL Setup

automatically trained model. We then show that fixing such alignments leads to reduction of error in word alignment, as measured by AER [Fraser and Marcu, 2007a]. We compare this with a baseline where links are selected at random for manual correction. To run our experiments iteratively, we automate the setup by using a parallel corpus for which the gold-standard human alignment is already available. We select the Chinese-English language pair, where we have access to 21,863 sentence pairs along with complete manual alignment.

Results

We first automatically align the Cn-En corpus using GIZA++ [Och and Ney, 2003]. We then use the learned model in running our link selection algorithm over the entire corpus to determine the most uncertain links according to each active learning strategy. The links are then looked up in the gold-standard human alignment database and corrected. In case a link is not present in the gold-standard data, we introduce a NULL alignment, else we propose the alignment as given in the gold standard. We select the partial alignment as a set of alignment links and provide it to our semi-supervised word aligner. We plot performance curves as number of links used in each iteration vs. the overall reduction of AER on the
Figure 3.4: Performance of active sampling strategies for link selection

corpus.

Query by committee performs worse than random indicating that two alignments differing in direction are not sufficient in deciding for uncertainty. We will be exploring alternative formulations to this strategy. We observe that confidence based metrics perform significantly better than the baseline. From the scatter plots in Figure 3.4 \(^1\) we can say that using our best selection strategy one achieves similar performance to the baseline, but at a much lower cost of elicitation assuming cost per link is uniform.

We also perform end-to-end machine translation experiments to show that our improvement of alignment quality leads to an improvement of translation scores. For this experiment, we train a standard phrase-based SMT system [Koehn et al., 2007] over the entire parallel corpus. We tune on the development set from MT-Eval 2004 dataset and test on a subset of the MT-Eval 2004 test set consisting of 631 sentences. We first obtain

\(^1\)X axis has number of links elicited on a log-scale
3.6 Proposed Work

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>18.82</td>
<td>42.70</td>
</tr>
<tr>
<td>Human Alignment</td>
<td>19.96</td>
<td>44.22</td>
</tr>
<tr>
<td>Active Selection 20%</td>
<td>19.34</td>
<td>43.25</td>
</tr>
</tbody>
</table>

Table 3.1: Alignment quality and Translation Quality

the baseline score where no manual alignment was used. We also train a configuration using gold standard manual alignment data for the parallel corpus. This is the maximum translation accuracy that we can achieve by any link selection algorithm. We now take the best link selection criteria, which is the confidence based method and train a system by only selecting 20% of all the links. We observe that at this point we have reduced the AER from 37.09 AER to 26.57 AER. The translation accuracy as measured by BLEU [Papineni et al., 2002] and METEOR [Lavie and Agarwal, 2007] also shows improvement over baseline and approaches gold standard quality. Therefore we achieve 45% of the possible improvement by only using 20% elicitation effort.

3.6 Proposed Work

3.6.1 Ensemble techniques for sentence selection

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. In this thesis we propose a switching ensemble approach that switches from one query strategy to another. Our approach is inspired from the DUAL ensemble 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.

Let us consider the dden-div approach in more detail. It has two components for scoring a sentence $S$, a density component $d(s)$ and a diversity component $u(s)$ as mentioned in Equation 3.5. The dden-div 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.
In the initial phases of evolution of an MT system, there is very little or no labeled data, hence every sentence is highly diverse. The dden-div approach 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, dden-div may score sentences with moderate density higher than the sentences with high diversity, there by making this criterion suboptimal. We propose to address this weakness with a new ensemble approach called 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.
Chapter 4

Multi Annotation Active Learning

Where as in chapter 3 we explored building general purpose MT systems by successfully eliciting a parallel corpora via active learning, in this chapter we argue that the traditional framework of active learning for eliciting a single kind of annotation needs to be extended to work with multiple types of annotation. We explore this in the context of transfer learning for machine translation, in particular building multi-domain translation systems.

4.1 Introduction

The traditional setup for active learning assumes a single model for which the selection of data is optimized. Given the nature of supervised learning tasks of the current day, we can say that this is a strong assumption. For instance recent work in parsing shows that named entity recognition task and parsing tasks can be combined to improve overall performance of a parser [Finkel and Manning, 2010]. Such setups are increasingly becoming popular in NLP and are called joint learning or multitask learning [Caruana, 1997]. Multitask learning is a case of transfer learning [Pan and Yang, 2010] and can be seen as transfer of knowledge from one task to the other in order to bridge difference in assumptions across both the tasks and in the process reduce the effort or increase the performance of either or both the tasks. When multiple tasks are being trained together in novel setups like joint-learning, multitask learning, there is also a need for new active learning strategies that can elicit informative samples that improve the performance for these learning scenarios. We call this ‘multi annotation active learning’. Recent work in active learning for multitask learning proposes an extension of the single-sided active elicitation task to a multi-task scenario, where data elicitation is performed for two or more independent tasks at the same time [Reichart et al., 2008a]. [Settles et al., 2008] propose elicitation of annotations for image segmentation under a multi-instance learning framework.
We first discuss two multitask learning scenarios to define the scope in which this thesis addresses multitask learning. We then propose active learning strategies for the same to jointly reduce the effort involved in both the tasks. Our learning strategies can be extended to other learning problems that fit these scenarios, but in this thesis we only experiment their applicability to MT.

• **Scenario 1:** In this setup, the two learning problems differ in both feature space, input and output spaces and so there is little incentive in training these tasks together. However, the output of one task constrains the input space of the other.

• **Scenario 2:** In this scenario we assume that the different tasks share the labeled and unlabeled data and also have an overlap in the feature space. Therefore there is an incentive to train these systems together making use of the available labeled data from the individual tasks.

Active learning for MT has not yet been explored in its full potential. Much of the literature has explored one task - optimising the selection of monolingual sentences for translation elicitation [Ambati et al., 2010, Haffari et al., 2009]. In our work we independently explored a second annotation, completely independent from the first task, in the form of word alignment. However, an MT system has a variety of annotations that result in potential improvement of translation quality. In context of an SMT system, some types of annotations that are of value are - translation of individual words or phrases, named entity tagging, syntactic annotation of an existing or a new sentence, post-editing of system-generated translations, evaluation of system generated output among others.

In this chapter we will first design two different machine translation setups that extend from the multitask learning scenarios discussed above, where we will explore the benefit of multiple annotations. We will then devise multi annotation active learning strategies for each of these two scenarios.

• **Focused domain MT** is the problem of building a domain specific translation system by first acquiring focused unlabeled data using a text classifier built for the domain and then using this domain membership information in obtaining translations to create a translation system. This is a clear instance of scenario 1 of multitask learning, where the output of the first task, constraints the input for the second task.

• **Multi domain MT** is the task of building domain-specific MT systems for the multiple domains that we are interested in. We will show that this task fits the scenario 2 of multitask learning, as the labeled data is shared across multiple domains and the goal is to improve all the systems simultaneously.
4.2 Focussed Domain Machine Translation

Very often we are interested in building low-resource language translation systems that address a particular immediate need and are hence domain specific. For example, the recent earthquake in Haiti, required translation of Haitian Creole into English in order to help the relief volunteers communicate with the local people and also disseminate information. The need was to quickly build MT systems that could address medical, travel, emergency domains.

Domain specificity is also required for translation in majority languages. With the explosion of Internet, current translation service providers have a need to cater to the demands of users by adapting to new trends in data, new unseen domains and even genres like social media. It is unlikely to find parallel corpora for every such scenario, and creation of such resources needs more resources than we can expend in time and money.

Our approach to building a focussed domain MT system is to first obtain domain specific unlabeled data using sentence classification. Our framework is shown in figure 4.1. Our hypothesis is that the task of active sentence selection can be improved by using domain membership knowledge from a first task of sentence categorization. We improve the sentence categorization by eliciting a second annotation in the form of domain tagging.

Figure 4.1: Focussed Domain MT
4.2.1 Task 1: Sentence Classification

Given a corpus $U$ consisting of sentences from a set of domains $D$, our task is to take as input a source sentence $s$ and predict whether it belongs to a particular domain $D = d$ or not. This task is very similar in nature to document classification, but with sparse text. Support vector machines (SVM) have proven successful for the document classification task. SVMs have a nice property that make it a preferable binary classifier when classifying text [Joachims, 2001]. We can also use the SVM to output a posterior probability for the classification as $P(D = d|s)$.

We use SVMs for training our sentence classification. We frame this as a binary class classification problem. The performance of the classifier depends to a large extent on the features used. We tried using higher order features in the form of bigrams and trigrams and observe that unigrams are a better tradeoff given the sparsity of features at sentence level. Motivated by document classification, we also tried more sophisticated features like term-frequency and inverse sentence frequency, and observe very minor improvements over binary features. So we stick to a very basic form of the features and use unigrams in the sentence as binary features.

4.2.2 Task 2: Sentence Translation

The second task is to actively select sentences from the unlabeled data for translation by an expert. Here, we will be selecting sentences from a monolingual dataset that is specific to a domain as classified by the sentence classifier above.

4.2.3 Query strategy for multiple annotations

Our active learning algorithm for this particular task takes the following approach. We first improve performance of the sentence classification task by actively selecting training sentences. We use a density based diversity approach in order to favor sentences that provide a larger coverage of the feature space, unigrams in our case. The scoring function $tc\_score(s)$ is defined below, where $Voc(s)$ is defined as he vocabulary of sentence $s$ in the unlabeled, mixed domain corpus $U$.

$$
tc\_score(s) = \sum_{s=1}^{[U]} sim(s, s') \cdot \frac{1}{|U|} \tag{4.1}
$$

$$
sim(s, s') = ||(Voc(s) \cap Voc(s'))|| \tag{4.2}
$$

The improved classifier model is used to classify the domain specific untranslated data.
4.2 Focussed Domain Machine Translation

as $U_d$, where $d$ is the particular domain that we are interested in. The classifier posterior probability distribution, $P(D = d|s)$, is also used also in the second task to actively select sentences for translation. Our diminishing density weighted sentence selection algorithm discussed in chapter 3 is weighted with this probability from the classifier as below, where $U_d$ is the unlabeled domain specific data and $L_d$ is the domain specific parallel data and $Phrases(s)$ is the same as before, the phrases in a sentence up to a certain length $n = 3$.

$$s^* = \arg \max_s P(D = d|s)V(s|U_d, L_d)$$  \hspace{1cm} (4.3)

$$V(s|L) = \sum_{x}^{Phrases(s)} P(x|U_d) * e^{-\lambda \text{count}(x|L_d)} \ |Phrases(s)|$$  \hspace{1cm} (4.4)

$$\|Phrases(s)\|$$  \hspace{1cm} (4.5)

The key step is to decide when to switch from eliciting one kind of annotation to the other. We compare a moving average of the performance to observe when the performance gains from further training sentences starts to plateau as observed on a held out dataset of domain tagged set of sentences. In our case, the sentence classification task is evaluated using accuracy of classification and therefore we compute the slope of this curve plotted across iterations to observe the diminishing effect of returns. We decide to stop when the difference in slopes across different iterations drops below $\delta = 0.5$ accuracy units.

4.2.4 Experiments

We perform two sets of experiments. Firstly, we are interested in learning the accuracy of the sentence classification model and how the active learning component affects its performance. Then we also experiment to see the effect of the first task on the overall translation quality.

Active improvement of sentence categorization

We perform this experiment with simulated data, where we have domain information for the dataset apriori. We selected Spanish-English as our language pair, where we have data from two different domains - political and travel. The political data, consisting of 240K sentences is a subset of the Europarl data and the travel data consisting of 121K sentences was released as part of the BTEC [Takezawa et al., 2002] corpus from IWSLT. The unlabeled data was created by combining the two datasets and randomizing the sentences. In order to simulate the annotation phase of active learning we reveal the domain information to create training data for the text classifier. We also held out a 2000 sentence test set prepared by drawing 1000 sentences each from both the domains. We compare our active
selection strategy to a random baseline, where the sentences for training the classifier were drawn randomly from the unlabeled data. Figure 4.2 shows the accuracy curves for the sentence classifier with the y-axis showing accuracy on the 2000 sentence test set and the x-axis showing the number of training samples drawn from the unlabeled data. The active selection outperforms a strong random baseline in a significant manner. Observing the graph, we note that while the random selection strategy requires 2000 sentences to reach 78% accuracy on the test set, the active sampling approach reaches the same accuracy with only 600 sentences, i.e. a 75% reduction in the amount of domain tagged training data.

![Figure 4.2: Active learning for sentence classification task](image)

**Translation Performance**

We then ran experiments for building a focused domain translation system. We wanted to understand the effect of sentence categorization quality on the active selection strategy. In other words, how well does the classification accuracy transfer into translation accuracy or how much does noise effect the AL strategy and resulting MT system.

We test it on the Spanish-English data discussed above and perform end-to-end translation experiments. The evaluation setup is similar to the sentence selection evaluation setup in chapter 3. Where as for the sentence selection task we had in-domain unlabeled monolingual data $U$ to choose sentences from, here we use the $U_d$ sentences tagged by the sentence classifier as being in travel domain. We also use a step function for the posterior distribution $P(D = d|s)$ and therefore every sentence tagged by the classifier is used in the sentence selection phase with uniform weight. We will relax this in our future experiments.
We are interested in building a travel domain system and so the ideal scenario is where we have access to the entire BTEC data tagged as belonging to travel domain (InDomain). We also have a baseline where we do not have an extra categorization phase and pass the mixed domain corpus to the active sentence translation phase (WildCorpus). We also tried two points on the accuracy curves for the first task where we switch between the two tasks. 'SwitchA' curve corresponds to switching at an accuracy level of 75% and the second 'SwitchB' corresponds to when the curve starts to plateau under the criteria discussed before. As expected the best results are observed in the 'InDomain' curve, where there was zero noise in the data. However, in reality such quality can only be achieved by marking all the sentences by a human which is very expensive. We show that using our two-stage active learning strategy we do much better than the baseline that was trained on the wild corpus. We also approach the best case scenario by automatically obtaining domain information for a very small portion of the MT training data.

![Figure 4.3: Active learning for sentence classification task](image)

4.3 Proposed Work: Focused Domain MT

Our experiments have shown that we can improve the quality of a domain specific MT system in the multitasking scenario. We propose work along the following multiple lines:
4.3.1 Selection strategies for sentence classifier

In our preliminary work we have tried a density based selection of sentences for improving our sentence classifier. While this has the advantage that we can use this with any classification model, we can do better with information of the model and its uncertainty. Active learning for document classification task has been explored well for a variety of classification models like logistic regression [Andrew et al.], SVM [Tong and Koller, 2002] etc. We will explore similar query selection strategies for selecting samples for the sentence classifier.

Uncertainty sampling:

We will also experiment by sampling sentences where the model is highly uncertain about the prediction. In order to be able to do this, we need the classifier to provide normalized classification scores for all the output classes. This information can be obtained by proper calibration of the posterior distribution of the SVM classifier.

\[ s^* = \arg \min_s P(D = d|s) \]

Margin sampling:

Sampling instances around the decision boundary of the classifier helps refinement of the model quicker. Given a classifier with posterior distribution over classes for an instance \( P(D = d/s) \), the margin based strategy is framed as below, where \( d_1 \) is the best prediction for the domain and \( d_2 \) is the second best prediction under the model. Intuitively, we select sentences where the model only selects a narrow winner.

\[ s^* = \arg \min_s P(D = d_1|s) - P(D = d_2|s) \]

4.3.2 Sentence vs. Document Classification

Sentence classification although similar to document classification, is not a natural task for humans to complete. Firstly it is very difficult to identify the domain of a sentence when taken out of context, making the tags as such less reliable when run in a real-world scenario. Secondly, most data on the web is classified logically into larger chunks like paragraphs or pages where one can weakly assume that all sentences in the chunk belong to a single domain. We will run experiments to substitute sentence classification with document
classification and re-run the experiments. We do not start with document classification task, as we do not know of parallel corpora for MT that are categorized at document level in both languages.

### 4.3.3 Cost Models

We will frame and study the multi annotation active learning under a cost-sensitive framework. More formally, consider multiple annotation tasks that elicit different types of annotation $i \in 1...n$ from the unlabeled dataset $U$. The goal of multi-annotation active learning is to select the optimal set of instances for each annotation $i$. Let the number of instances elicited for each annotation per iteration be $k_1...k_i...k_n$ and the cost models for be $c_1...c_i...c_n$. However, instead of optimizing the number of instances selected, we optimize the selection under a provided budget $B$ per iteration of the active learning algorithm, and $|U|$ is the size of the total unlabeled data set. Therefore our cost based formulation will be as seen below and therefore our evaluation of the active learning will also be geared towards reducing overall cost across the multiple annotation tasks.

$$B_i = c_i \times k_i$$

$$\text{where} \sum_{i}^{n} B_i \leq B, \forall k_i \leq |U|$$

The cost-sensitive framework will make it possible to run interesting experiments.

- **Real-world costs:** We can obtain real-world cost information from online platforms like Mechanical Turk, and plug them into our setup to study the effect of multiple annotations.
- **Skewed costs:** We can simulate scenarios under extreme parametrization of the cost models, and study the benefit of using multiple annotations. For instance one interesting configuration is where the document/sentence classification is 10 times lower than the actual translation task.

### 4.3.4 Switching criteria

As we observed in our preliminary experiments selecting an optimal point for switching from task1 to task2 affects the overall performance of the translation. We currently use an average computed over a window to switch from one task to another. We will explore more advanced strategies to decide the switching point.
We currently have access to three domains - politics (Europarl, UN datasets), medical (BTEC) and travel (BTEC) for Spanish-English. We will extend our experimentation to these domains. Finally, we will extend our work which was currently performed on Spanish-English to other languages and also new domains.

4.4 Proposed Work: Multi-Domain Machine Translation

We will first design the setup for building multiple domain translation systems under scenario 2 of multitask learning and propose some ideas for improving it by multi annotation active learning.

4.4.1 Problem setup

Given unlabeled data sets for multiple domains and a fixed budget to build translation systems for multiple domains, we propose to build multi-domain translation systems simultaneously. As shown in Figure 4.4, we have a set of domains $D$ each with monolingual unlabeled data $U_i$, where $i$ is an index over the set of domains $D$. The active learning module will then draw labeled sentences to be annotated by an expert translator. We assume that the translator is also provided with the domain information $D_i$ of the sentence $s$ and therefore he provides a translation $t$ within the context of the chosen domain. The labeled data is now available to train all the individual MT systems, which in turn will be tuned and tested on their respective domain specific development and test sets.

Devising active learning schemes for multitask learning in this scenario requires optimally eliciting multiple annotations for the individual MT system building tasks, where providing translation for a domain specific sentence can be treated as a different kind of annotation. Given multiple kinds of annotations, a natural question to ask is when to seek which annotation for what instance. We propose multiple active learning strategies for this scenario such a case to reduce the annotation effort and cumulatively improve all the MT systems simultaneously.

4.4.2 Query strategies

Maximum Local Utility

This is a greedy approach to always select the best scoring sentence across domains according to a single-domain active learning strategy $V(s)$. We will select our best performing single-domain active learning strategy from Chapter 3. $U_i$ and $L_i$ are the domain specific unlabeled and labeled data respectively. We also assume a soft domain membership for each sentence.
4.4 Proposed Work: Multi-Domain Machine Translation

Figure 4.4: Multi-domain MT using Active transfer learning

provided by $P(D_i|s)$ and a uniform cost of translation $C_i$, within the domain. The parameter $\lambda_i$ is used to capture importance of each domain and can be initialized based on human input, or dynamic monitoring usage of the MT system over time. In a realworld scenario, higher the usage, the value of $\lambda_i$ is higher for that domain indicating that focussing in this domain is more beneficial.

$$\left(s^*, i^*\right) = \arg \max_{s \in U_i} \lambda_i \frac{P(D_i|s)V(s|U_i, L_i)}{C_i}$$  \hspace{1cm} (4.6)

Maximum Global Utility

This is another greedy approach to always select the best scoring sentence across domains. However we will score all the sentences in one domain under the value functions of all the other domains, thereby giving a global best sentence that is useful to most domains.

$$\left(s^*, i^*\right) = \arg \max_{s \in U_i} \sum_i \lambda_i \frac{P(D_i|s)V(s|U_i, L_i)}{C_i}$$  \hspace{1cm} (4.7)
Expected accuracy

One disadvantage with the above approaches is that they assume that the value of each sentence according to a learning strategy is directly correlated to the end translation quality for the corresponding domain. In general, the costs for annotation of each of the sentence are going to be non-uniform and so are the benefits to an MT system.

We will compute the delta improvement of translation quality in each domain across iterations and then estimate the contribution of a particular domain to the overall score. We will use this as an approximation to the expected accuracy contribution and select sentences that maximize the same.

4.4.3 Evaluation

We will evaluate the above query strategies and compare them against multiple baseline approaches. A first baseline is to combine all the individual unlabeled data and run our active sentence selection algorithm as designed for a single MT system. Although one would imagine this baseline to be weak as we do not use any domain information, our expectation is that it will be a strong baseline as our selection strategies will extract the common parts in the combined dataset. A second baseline is to draw uniformly from unlabeled data of all the domains. This will ensure representation from all domains and we expect this to be a good baseline.

The success criteria in a multi-domain MT system should be a cumulative score indicative of the improvements in the individual MT systems as shown in Eq 4.8. We will use BLEU or METEOR automatic metrics to evaluate the performance of the individual MT system on their respective test sets. \( \lambda_i \) is same as before, indicating the importance of the domain.

\[
multieval = \sum_i \lambda_i BLEU_i
\]  

(4.8)

We will also study this under a cost-sensitive framework, where the cost of acquisition of translations is different for different domains, due to either availability of translators, difficulty of translation etc. Evaluation in such a case will also include an overall reduction of cost of building the translation systems.
Chapter 5

CrowdSourcing for Machine Translation

In all our experiments in Active Learning, we have so far made an implicit assumption of the oracle. We assume that the oracle is present, and provides with annotations of his expertise. However, for some learning tasks it may be difficult to find experts for annotation. In this chapter we will explore the dimension of the annotator. More importantly, we will study the ‘expert vs. non-expert’ problem and propose crowdsourcing as a suitable alternative for eliciting annotation.

The data provided by an expert is of high quality and most desirable for improvement of a translation system. For some problems in NLP like parsing, syntactic annotation, there is no substitute for an expert. Translation on the other hand, is a slightly less demanding task, which can perhaps be completed by non-linguists or native speakers of the language, who have some familiarity with a second language. For example, to translate the English sentence “Is there a hotel near by?” into Chinese, we can make do with the availability of a native Chinese speaker who understands English. And to create a translation system for a travel domain, where we will need a few thousands of such English sentences translated into Chinese, we should be able to make use a non-linguist’s help. It is this aspect of translation that we want to explore. The following research questions will be pursued:

- How do we collect data from the crowd? What are some of the challenges involved in bringing humans in the loop of building translation systems?
- Our experiment with non-experts highlight the need for quality assurance of crowd data which is an onerous task if done manually. We will explore ways of automatically exploiting overlapping labels from multiple non-experts to improve quality.
- Eliciting data from a large numer of non-experts may improve quality, but the associ-
ated cost will quickly become prohibitive. We will explore the cost vs. quality trade-off and devise techniques to elicit high quality data at reduced costs.

5.1 Introduction

Crowd-sourcing is the process of farming out tasks to a large number of users over the web. These tasks are typically performed by a resident employee or a contractor with a specific area of expertise. With crowd-sourcing such tasks are requested from a crowd, which is a set of non-experts. Recently, crowd-sourcing has become popular as human computing, where tasks that are challenging, difficult or time-consuming for computers are passed to human crowds over the web. These tasks broadly belong to the language or vision community, where for a number of tasks it is still impossible for computers, but only requires a few seconds for a human to complete. For example, identifying a person in a photograph, tagging a video for a particular event, flagging an email for spam, identifying the sentiment of a written text, spotting characters in an image are still some of the challenge research problems to computers.

A few variations of the concept of crowd-sourcing exist, with the popular being the paid version of crowd-sourcing which is enabled by third-party platforms like Amazon Mechanical Turk, where each task is highly granular and is completed by a person.

Figure 5.1: Active Learning setup

In our work we use Amazon Mechanical Turk (MTurk) for crowd-sourcing. Amazon Mechanical Turk (Mturk) is an online marketplace that enables computer programs to coordinate with humans via crowd-sourcing. A typical workflow for data collection on Mturk is depicted in Figure 5.1. Requesters can pose tasks known as HITs (Human Intelligence Task), and workers, also known as turkers, can then browse among existing tasks and complete them for a payment provided by the requester. Payments are processed by their micro-payment system which is available both in dollar (USD) and Indian rupee (INR). In order to ensure quality, MTurk provides screening based on worker parameters like
5.2 Data Collection for MT

Crowd-sourcing compensates for the lack of experts with a large pool of expert/non-expert crowd. However, crowd-sourcing has thus far been explored in the context of eliciting annotations for a supervised classification task, typically monolingual in nature [Snow et al., 2008]. In this project we test the feasibility of eliciting parallel data for Machine Translation (MT) using Mechanical Turk (MTurk). MT poses an interesting challenge as we require turkers to have understanding/writing skills in both the languages.

Recent efforts in MT include feasibility studies for using crowd sourcing techniques for MT Evaluation; users are provided with translations from multiple systems and asked to select the correct one [Callison-Burch, 2009], [Zaidan and Callison-Burch, 2009]. One observation that [Callison-Burch, 2009] make is about the availability of bilingual speakers for annotation tasks in MT. They observe that it is relatively more difficult to find translators for low-resource languages like Urdu, Thai, etc. than it is to find for Chinese, Arabic, Spanish, etc. With the increasing pervasiveness of the Internet, and more and more people in the developing world gaining computer literacy, the situation should ameliorate.

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. We provide detailed instructions on the HIT for both completion of the task and its evaluation. 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.

5.2.1 Language Landscape of MTurk

We selected 3 language pairs which we explored in greater detail during the course of the project. The language pairs are Telugu-English, Urdu-English and Spanish-English. Ideally, we would like a turker who is native speaker of the language which he/she is translating into. But as we will see in our data analysis, in almost all the cases we observe that is not the case. We sampled 100 sentences for each language-pair and requested three translations for each sentence. The Spanish data was taken from BTEC [Takezawa et al., 2002] corpus, consisting of short sentences in the travel domain. Telugu data was taken from the sports and politics section of a regional newspaper. For Urdu, we used the NIST-Urdu Evaluation 2008 data. We report results in Table 5.1. The goal of the experiment was to study the task design and feasibility of switching translation direction.

The first batch of HITs were posted to collect translations into English. The ideal
<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Cost</th>
<th>#Days</th>
<th>#Turkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish-English</td>
<td>$0.01</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Telugu-English</td>
<td>$0.02</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Urdu-English</td>
<td>$0.03</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>English-Spanish</td>
<td>$0.01</td>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>English-Telugu</td>
<td>$0.02</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>English-Urdu</td>
<td>$0.03</td>
<td>2</td>
<td>21</td>
</tr>
</tbody>
</table>

Table 5.1: Sentence translation data

target population would be native speakers of English who also understand the source language. We noticed from manual inspection of the quality of translations that most of our translators were non-native speakers of English. This calls for adept and adequate methods for evaluating the translation quality. For example more than 50% of the Spanish-English tasks were completed in India, and in some cases a direct output of automatic translation services.

The second set of experiments were to test the effectiveness of translating out of English. The ideal target population for this task were native speakers of the target language who also understood English. Most participant turkers who provided Urdu and Telugu translations, were from India and USA and were non-native speakers of English. However, one problem with enabling this task was the writing system. Most turkers do not have the tools to create content in their native language. We used ‘Google Transliterate’ API \(^1\) to enable production of non-English content. This turned out to be an interesting HIT for the turkers, as they were excited to create their native language content. This is evident from the increased number of participant turkers. Manual inspection of translations revealed that this direction resulted in higher quality translations for both Urdu and Telugu and slightly lower quality for Spanish.

5.2.2 Challenges for Crowd-Sourcing and Machine Translation

Low Quality

Quality assurance of crowd sourced data is necessary due to two reasons. Firstly, there are gamers on the web who would could either trick the task for money and thus introduce random noise into the annotation. Problems like blank annotations, mis-spelling, copy-pasting of input are prevalent, but easy to identify. Turkers who do not understand the task but attempt it anyway are the more difficult ones to identify, but this is to be expected with non-experts. Secondly, annotators available on the web may mis-interpret the task.

\(^1\)http://www.google.com/transliterate/
description and could complete the task incorrectly.

**Turking Machines:**

We also have the problem of machines posing as turkers – ‘Turking machine’ problem. With the availability of online translation systems like Google translate, Yahoo translate (Babelfish) and Babylon, translation tasks on MTurk become easy targets to this problem. Turkers either use automatic scripts to get/post data from automatic MT systems, or make slight modifications to disguise the fact. This defeats the purpose of the task, as the resulting corpus would then be biased towards some existing automatic MT system. It is extremely important to keep gamers in check; not only do they pollute the quality of the crowd data, but their completion of a HIT means it becomes unavailable to genuine turkers who are willing to provide valuable translations. We, therefore, collect translations from existing automatic MT services and use them to match and block submissions from gamers. We rely on some gold-standard to identify genuine matches with automatic translation services.

**Output Space**

Due to the natural variability in style of turkers, there could be multiple different, but perfectly valid translations for a given sentence. Therefore it is difficult to match translation outputs from two turkers or even with gold standard data. We therefore need a fuzzy matching algorithm to account for lexical choices, synonymy, word ordering and morphological variations. This problem is similar to the task of automatic translation output evaluation and so we use METEOR [Lavie and Agarwal, 2007], an automatic MT evaluation metric for comparing two sentences. METEOR has an internal aligner that matches words in the sentences given and scores them separately based on whether the match was supported by synonymy, exact match or fuzzy match. The scores are then combined to provide a global matching score. If the score is above a threshold $\delta$, we treat the sentences to be equivalent translations of the source sentence. We can set the $\delta$ parameter to different values, based on what is acceptable to the application. In our experiments, we set $\delta = 0.7$. We did not choose BLEU scoring metric as it is more oriented towards exact matching and high precision, than towards robust matching for high recall.

5.3 Quality in Crowd

Quality control is a concern with an online crowd where the expertise of the turkers is unknown. We also notice from the datasets we receive that there exist a lot of consistently poor and noisy translators lacking the necessary minimal expertise. Therefore it is important
to not only cleanup the data to get rid of noisy annotations, but also to have an estimate of the quality of the data. These estimates of reliability can help us use the annotation accordingly in down-stream tasks. We propose the following two paradigms to improving quality when working with multiple non-experts.

5.3.1 Selective Approaches

Redundancy of translations for the input and computing majority consensus translation is agreed to be an effective solution to identify and prune low quality translation. We discuss in following section computation of majority vote using fuzzy matching. For a language pair like Urdu-English, we noticed a strange scenario, where the translations from two turkers were significantly worse in quality, but consistently matched each other, there by falsely boosting the majority vote. We suspect this to be a case of cheating, but this exposes a loop in majority voting which needs to be addressed, perhaps by also using gold standard data. The problem with these translators is that not only do they make the average cost of the task expensive, but they also make it difficult to estimate the quality of translations by skewing the distribution of the majority consensus translation.

In this section we discuss how we plan to use the technique of redundancy effectively to collect such reliability estimates. To ensure quality, we plan to request annotation for each input from multiple Turkers.

Annotation 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 or annotated similarly by two or more annotators, the more likely it is to be a correct annotation. In our preliminary work 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 annotators agree with each other, 23.8% two translators agree, and 55.1% there was no agreement between translators.

In case of translations, in our preliminary work we use exact matching to compare them. This is not robust to variations in spelling or other language phenomenon. 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. In future we will also be exploring other sophisticated methods for translator reliability estimation similar to [Donmez et al., 2009]. A more challenging task is to perform matching when there could be more than one semantically valid translations for a given sentence.
5.3 Quality in Crowd

Annotator Reliability Estimation

The above approach of seeking multiple annotations 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 translations \( T_k = \{t_{kj}\} \) that he/she translated, we estimate reliability based on translations from other users \( U_n = \{t_{nj}\} \) as shown in equation below.

\[
rel(W_k) = \frac{\sum_{t_j \in T_k} \sum_{n_i \in U} \alpha}{\|T_k\|}
\]

\[
\alpha = \begin{cases} 
1 & t_{kj} \equiv t_{nj} \\
0 & \text{otherwise}
\end{cases}
\]

Weighted Majority Voting

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.

5.3.2 Collective Approaches

Select-All

Translations from humans vary considerably. We observe that the same sentence can be translated into different meaning-preserving variations, and there is not a consensus in the word choice or the ordering of the sentence in the target language. As an example consider the translations provided by three turkers.

Spanish: Me alegra mucho que hayas podido venir

- I am so glad you could come
- I’m very happy you were able to come
• I am glad you could make it

We notice that all the translations are valid and yet there can not be a ‘majority translation’ by any fuzzy match. Such sentences are quite helpful to capture the paraphrasing phenomenon in MT that has proven to be useful [Callison-Burch et al., 2006]. We observe that this does occur quite often in our experiments.

Our approach is to use all the translations to build a translation model. We assume that the maximum likelihood training of the statistical phrase translation models would accenuate the good fragments of the translations and demote the less agreed parts of the translations. In case of good translations that are just paraphrases, the availability of alternatives at decoding time will ensure better coverage as well.

5.4 Cost Effective Elicitation

Although the approach of seeking multiple translations from turkers works great in accounting for natural variability of translators and reduces human error, it may not be a viable long-term strategy as it increases the cost of annotation. In this section, we first motivate the need for minimizing cost even in a low-cost setting like MTurk and then we discuss the application of two cost minimization approaches for data collection.

Cost Factor: Crowd-sourcing has been used to reduce the cost of annotation for NLP [Callison-Burch, 2009, Snow et al., 2008]. However, our experiments in MT have shown that this may not necessarily be the case. We found that at lower pay rates, it is difficult to find a sufficient number of annotators to complete even a single assignment of the task. For example, we could not find turkers to complete the translation of 100 Haitian-Creole sentences into English even after a period of 10 days. Haitian-Creole is spoken by a small population and it seems that only a very small portion of that was on MTurk. Understandably, the recent disaster in Haiti, may also have been a factor in the low activity for this language. For a few other languages pairs, while we could find a few turkers completing the task, the price had to be increased to attract any attention. We selected 100 sentences from each language pair (with length under 10 words) and posted on MTurk to collect translations. Table 5.2 shows the minimum cost at which we could start getting turkers to provide translations and the number of days they took to complete the task. This scenario highlights not only the need to obtain quality translation but also to do so within a limited budget constraint. MTurk has so far been a suppliers’ market, and tasks like translation show how one can only get a few turkers making it a demand-driven market.

Consistently low-quality annotators: From the datasets we receive, we also notice that there are consistently poor and noisy translators. The problem with these translators is that not only do they make the average cost of the task expensive, but they also make it difficult to estimate the quality of translations by skewing the distribution of the majority
### 5.4 Cost Effective Elicitation

The table below provides the cost per sentence, completion days, and number of turkers for various language pairs.

<table>
<thead>
<tr>
<th>Pair</th>
<th>Cost/sen</th>
<th>Days</th>
<th>#Turkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish-Eng</td>
<td>$0.01</td>
<td>1</td>
<td>101</td>
</tr>
<tr>
<td>Telugu-Eng</td>
<td>$0.02</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>Eng-Creole</td>
<td>$0.06</td>
<td>-</td>
<td>6</td>
</tr>
<tr>
<td>Urdu-Eng</td>
<td>$0.03</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Chinese-Eng</td>
<td>$0.02</td>
<td>1</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 5.2: Cost vs. Completion for Language pairs

consensus translation.

#### 5.4.1 Exploration vs. Exploitation

Given that low-quality translators can be dropped without deteriorating the quality, we propose a strategy to drop translators and reduce the total cost of elicitation. The general approach is to identify effective turkers early on (exploration) and get them to do more tasks (exploitation). Typically a gold standard data $G$ can be used to compute this. Consider we have a set of tasks $T$ and set of workers $W$, and $O \subseteq W$ is a set of workers who are suitable to complete $T$, and $|G| << |T|$. As a general rule, the exploration vs. exploitation approach is used only when $- (|G| * |O| + |T|) << (|T| * |W|)$, where $|.|$ represents the size of each set. We observe that for language-pairs where only a few translators are available, this approach may not be justified.

Using the exploration phase we identify reliable translators as early as possible. If the translations provided by the worker significantly deviate from the gold standard, we can suspect incompetency of the worker for the task. In MTurk, there is no provision to direct HITs to a specific turker, but we can attract desired turkers by providing incentives in the form of a bonus. We can also maintain a local log of preferred turkers for our task and favor the annotation from a preferred turker, when the quality is higher than that of others who attempt the same task.

#### Relation to Interval Estimate Threshold

[Donmez et al., 2009] propose an approach called ‘IEThresh’ to select, at every stage of annotation, the right annotator for annotating an instance. They estimate a confidence interval for the reliability of each annotator and filter out ones that are below a certain threshold. We implement their approach and adapt to the task of translation using fuzzy matching to identify the majority vote. If none of the translators are filtered out, this approach is equivalent to selecting the majority vote translation, which could produce high quality data, but at a high cost. This approach works better when a large number of data
points can be obtained for each translator in order to establish a good confidence interval for the initial reliability. Therefore we set a very low threshold on the confidence interval in the initial stages in order to not filter out too many translators. As we obtain a better estimate, the threshold can be made much stricter to select good translators only.

5.5 Experiments

We performed our experiments on the Spanish-English language pair on two batches. In each batch, we selected 1000 Spanish sentences that were crowd-sourced for translation via MTurk. Each sentence was presented to three different turkers for translation. The first batch of tasks were completed by a total of 71 turkers, to provide 3000 translations. This batch consisted of sentences that were less than 7 words. A total of 17 man hours was spent among these turkers. The second batch which consisted of sentences less than 12 words, was completed by 101 turkers. The total cost for both batches was within 100 USD. We evaluate all our experiments by scoring against a gold standard data set for each of these batches. Tables 5.3, 5.5 show results of quality maximization on both short and medium sentence respectively. Tables 5.4, 5.6 show results of cost reduction on the two batches.

5.5.1 Quality

We evaluate several different selection strategies on these two data sets. For each sentence in the batch, we apply our selection strategy to select one translation from among all the submitted translations for the sentence. We then compare the selected translations with the gold standard data and report results from METEOR [Lavie and Agarwal, 2007] and BLEU [Papineni et al., 2002] automatic metrics. Our baseline is a random selection strategy, where a translation was selected at random from among the multiple translations. We observe that all our selection strategies work similar to or better than the baseline. ‘Rel vote’ approach always prefers the output of that translator who agrees the most with his peers over the entire batch. Similarly ‘Maj-vote’ always prefers the consensus translation, while ‘W Maj-vote’ is the weighted majority voting approach discussed in Section ??. We observe that in the first batch of sentences, as they are short we find that the quality of translations were good, with a strong baseline performance. Our selection strategies perform better with weighted majority voting, providing a better selection. In batch 2 however, we find a drop in quality showing that there were low quality translations which skewed the majority vote approach. In such a scenario, also using worker reliability estimates works better over pure majority voting.
5.5 Experiments

<table>
<thead>
<tr>
<th>Strategy</th>
<th>BLEU</th>
<th>METEOR</th>
<th>Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>33.03</td>
<td>57.03</td>
<td>1000</td>
</tr>
<tr>
<td>Rel Vote</td>
<td>33.75</td>
<td>57.83</td>
<td>3000</td>
</tr>
<tr>
<td>Maj Vote</td>
<td>34.47</td>
<td>58.11</td>
<td>3000</td>
</tr>
<tr>
<td>W-Maj Vote</td>
<td>34.98</td>
<td>58.75</td>
<td>3000</td>
</tr>
</tbody>
</table>

Table 5.3: Evaluation of Quality maximization with short sentences

<table>
<thead>
<tr>
<th>Cost Strategy</th>
<th>BLEU</th>
<th>METEOR</th>
<th>Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp-Exp</td>
<td>34.15</td>
<td>57.76</td>
<td>2139</td>
</tr>
<tr>
<td>IEThresh</td>
<td>33.81</td>
<td>57.44</td>
<td>1870</td>
</tr>
</tbody>
</table>

Table 5.4: Evaluation of Cost minimization with short sentences

5.5.2 Cost

Assuming the cost of annotation of each input is uniform and the cost of eliciting a label from the annotator is uniform, we can equate cost to the number of queries or tasks posted for each strategy. As we cannot direct HITs at a particular turker on MTurk, we perform these experiments offline on the data collected. We experiment with exp-exp strategy by not requesting multiple translations when we have already received a translation from an explored worker. An explored or tested worker is one who has participated in at least 20 tasks and has a reliability score of 0.5 or above.

By doing so, we reduce the number of queries, but we risk relying upon a turker who only does well initially. IEThresh approach which relaxes this by continuously checking for performance as a confidence interval, was tested in a similar manner. We observe that, for low-resource language translation, where number of translators available are less, the Exp-Exp strategy works better than IEThresh [Donmez et al., 2009] by balancing the number of queries and quality.

5.5.3 Task based Evaluation

We also conduct end to end MT experiments in order to evaluate the effectiveness of our selection strategies in producing high quality parallel corpus using crowd data. We start with the first 1000 sentences from Spanish-English BTEC corpus that produces a BLEU score of 10.64 when tuned and tested on the IWSLT04 data sets. We create different versions of parallel corpora by using our selection strategies on the crowd translated short sentences. We then train and test an MT system on each of the parallel corpora versions separately. In Table 5.7 we show translation results as scored by BLEU. We only compare our best performing selection strategies. Two main observations are that, selection strategies that
Table 5.5: Evaluation of Quality maximization with medium sentences

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>METEOR</th>
<th>#Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>25.82</td>
<td>54.4</td>
<td>1000</td>
</tr>
<tr>
<td>Rel Vote</td>
<td>31.9</td>
<td>59.14</td>
<td>3000</td>
</tr>
<tr>
<td>Maj Vote</td>
<td>27.32</td>
<td>55.69</td>
<td>3000</td>
</tr>
</tbody>
</table>

Table 5.6: Evaluation of Cost minimization with medium sentences

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>METEOR</th>
<th>#Queries</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exp-Exp</td>
<td>28.93</td>
<td>57.01</td>
<td>2278</td>
</tr>
<tr>
<td>IEThresh</td>
<td>27.12</td>
<td>54.14</td>
<td>1956</td>
</tr>
</tbody>
</table>

explicitly model reliability of workers based on agreement, in fact produce good quality translation systems that mimic systems trained on expert data. As observed in previous section, synthesis also produces better quality data that improves final translation quality. A surprising result in our experiments is when we use all the human data with multiple translations for each sentence, we outperform the results from expert data. This could be due to the paraphrases in the non-expert translation, which we will explore further as part future work.

5.6 Proposed Work

Synthesize for Quality

Most translations produced by humans have a variety of errors as can be seen in the following example. We also notice that while there is no evidently better translation among the multiple translations above, a meaningful and complete translation can be synthesized from them. Spanish: lo tomar desde la parada de taxis

- i’ll take it from the taxi stop
- i’ll take it from the taxi rank
- i will take it from the taxi rank

Motivated by the above observations, we propose a synthesis approach for combining the good parts from all the translations instead of selecting the best one. This approach is inspired by the concept of multi-engine machine translation system (MEMT) [Nirenburg and Frederking, 1994] where a better output is synthesized from multiple MT system outputs. As our subjects here are humans instead of machines, we can treat this as a synthesis of human
translations. Traditional approaches to MEMT have looked at n-gram agreement features among candidate hypotheses [Heafield et al., 2009] in conjunction with other features, stronger language models. MEMT approaches assume systems to be non-adversarial and to produce their best output on each input instance. However due to low-quality of translators we propose to make use of reliability estimation techniques discussed above to bias the combination strategies. We will also be combining the reliability estimates computed from overlapping translations to initialize the multi-engine system for better performance. We believe that this can overcome the problem of consistently low performing and adversarial translators.

5.6.1 New setups for Collaborative Translation

Crowd-sourcing is becoming popular with researchers for cost-effective elicitation of annotations. However, quality of crowd data is a common concern in crowd-sourcing approaches to data collection. We will investigate collaborative methods for translation, where the participants are working with the output produced by each other. We believe that this setup may reduce the overall effort and improve the quality by reducing cheating and incorrect translations. We will then compare the effectiveness of this ‘wiki-style’ collaborative translation output to the output from paid crowd-sourcing model on Amazon Mechanical Turk, by evaluating how well it agrees with gold-standard expert translations.

Table 5.7: Automatic MT Evaluation using BLEU

<table>
<thead>
<tr>
<th>System</th>
<th>1k sens</th>
<th>+Batch 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>10.64</td>
<td>16.43</td>
</tr>
<tr>
<td>Maj Vote</td>
<td>10.64</td>
<td>18.92</td>
</tr>
<tr>
<td>W-Maj Vote</td>
<td>10.64</td>
<td>19.20</td>
</tr>
<tr>
<td>All</td>
<td>10.64</td>
<td>19.62</td>
</tr>
<tr>
<td>Expert</td>
<td>10.64</td>
<td>20.12</td>
</tr>
</tbody>
</table>
Chapter 6

Summary and Timeline

6.1 Summary of Work Done

6.1.1 Active Learning for MT

• We proposed active learning strategies for the sentence selection task. We show that our approach of prioritizing monolingual sentences leads to faster improvement of MT systems. We experimented with multiple languages including Spanish-English, Japanese-English and Urdu-English and show improvement across the board.

• We devised multiple learning strategies for eliciting alignment links that need manual attention for improvement of a semi-supervised word alignment model. We showed significant reduction of alignment error over strong baselines for both Chinese-English and Arabic-English language pairs.

6.1.2 Active Transfer Learning

• We formulated new multitask learning setups for machine translation. In particular we have identified two scenarios for machine translation where active learning and multitask learning can help.

• We showed improvement in building focussed domain MT systems by first formulating it as a combination of sentence classification and sentence translation tasks. We show improvement of both these tasks by active learning. A preliminary combination of these tasks via a switching strategy from one task to another shows significant improvement in the quality of a focussed domain MT system for Spanish-English.
6.1.3 Crowdsourcing for MT

- We have shown that translation data can be collected in a crowdsourced fashion for building an MT system.

- We proposed and experimented with metrics to select high quality data for training and MT system. Our results from Spanish-English language pair show that with effective methods we can build an MT system that approaches the quality of expert translation.

- We observed that cost can be an important factor for low-resource translation and proposed an algorithm for minimizing the cost of data collection.

6.2 Summary of Proposed Work

6.2.1 Active Learning for MT

- We will experiment with our switching ensemble technique for effectively switching between active learning strategies for different operating ranges in the evolution of an MT system. We will experiment with the same for multiple languages.

- We will also try the active selection strategies in a large-scale MT scenario to see how they work when an MT system already has a lot of existing parallel data.

6.2.2 Active Transfer Learning

- We will implement active selection strategies for jointly selecting instances for the two tasks in focused domain MT. We will also experiment this with more domains and languages (Scenario 1).

- We will implement sentence selection strategies for multi-domain MT and relate it to the task of multitask learning. We will show improvements over strong baselines of random, round-robin selection.

6.2.3 Crowdsourcing for MT

- We propose to design and experiment multi-engine MT approaches for improving quality of crowd data. Synthesis of a new translation using multiple translations from the crowd will also involve effectively biasing the combination towards reliable annotators.
• We are interested in new approaches to crowd-sourcing under a wikipedia style collaborative framework, with a goal of improving overall quality of translation. The approach will be evaluated against expert produced translations.

• We will also run more experiments for end-to-end active crowd translation using our ACT framework for atleast one new language pair

6.3 Schedule

• Nov- Dec 2010 (2 months)
  – Research:
    1. Joint selection of annotation and instance for focussed domain MT (Chapter 4)
    2. Implement proposed query strategies for the multi-domain MT (Chapter 4)
  – Writing: ACL 2010 submission

• Jan-Feb 2011
  – Research:
    1. Collaborative crowdsourcing approaches for translation (Chapter 5)
  – Writing: CHI workshop, Active learning workshop submission

• Mar-May 2011
  – Research:
    1. Switching ensemble involving data-driven to model-driven approaches for sentence selection (Chapter 3)
    2. More experiments on multi annotation active learning for new languages and domains (Chapter 4)
    3. Sentence selection for understanding structural and semantic divergences between multiple languages (Chapter 3)
  – Writing: EMNLP 2011

• Jun-Aug 2011
  – Research:
    1. Large-scale MTurk data collection and MT experiments for multiple languages under a budget motivated setup with real costs (Chapter 5)
2. Experiment synthesis and select-k approaches of crowd translations for data quality (Chapter 5)
   – Writing: Thesis writing, Job Search

- Sep-Nov 2011
  – Research:
    1. Deploy ACT to build an unseen and minority language setup
    2. Wrap-up experiments
  – Writing: Thesis write-up, Job Search

- December 2011 : Defense
Bibliography


I. Andrew, Schein, and Lyle H. Ungar. Active learning for multi-class logistic regression. 48


Pinar Donmez, Jaime G. Carbonell, and Jeff Schneider. Efficiently learning the accuracy of labeling sources for selective sampling. In KDD, pages 259–268, 2009. 23, 58, 61, 63

Pinar Donmez, Jaime Carbonell, and Jeff Schneider. A probabilistic framework to learn from multiple annotators with time-varying accuracy. In To appear in Proceedings of the SIAM Conference on Data Mining, 2010. 23


21, 25


