Active Learning with Multiple Annotations for Comparable Data Classification Task

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

Abstract

Supervised learning algorithms for identifying comparable sentence pairs from a dominantly non-parallel corpora require resources for computing feature functions as well as training the classifier. In this paper we propose active learning techniques for addressing the problem of building comparable data for low-resource languages. In particular we propose strategies to elicit two kinds of annotations from comparable sentence pairs: class label assignment and parallel segment extraction. We also propose an active learning strategy for these two annotations that performs significantly better than when sampling for either of the annotations independently.

1 Introduction

The state-of-the-art Machine Translation (MT) systems are statistical, requiring large amounts of parallel corpora. Such corpora needs to be carefully created by language experts or speakers, which makes building MT systems feasible only for those language pairs with sufficient public interest or financial support. With the increasing rate of social media creation and the quick growth of web media in languages other than English makes it relevant for language research community to explore the feasibility of Internet as a source for parallel data. (Resnik and Smith, 2003) show that parallel corpora for a variety of languages can be harvested on the Internet. It is to be observed that a major portion of the multilingual web documents are created independent of one another and so are only mildly parallel at the document level.

There are multiple challenges in building comparable corpora for consumption by the MT systems. The first challenge is to identify the parallelism between documents of different languages which has been reliably done using cross lingual information retrieval techniques. Once we have identified a subset of documents that are potentially parallel, the second challenge is to identify comparable sentence pairs. This is an interesting challenge as the availability of completely parallel sentences on the internet is quite low in most language-pairs, but one can observe very few comparable sentences among comparable documents for a given language-pair. Our work tries to address this problem by posing the identification of comparable sentences from comparable data as a supervised classification problem. Unlike earlier research (Munteanu and Marcu, 2005) where the authors try to identify parallel sentences among a pool of comparable documents, we try to first identify comparable sentences in a pool with dominantly non-parallel sentences. We then build a supervised classifier that learns from user annotations for comparable corpora identification. Training such a classifier requires reliably annotated data that may be unavailable for low-resource language pairs. Involving a human expert to perform such annotations is expensive for low-resource languages and so we propose active learning as a suitable technique to reduce the labeling effort.

There is yet one other issue that needs to be solved in order for our classification based approach to work for truly low-resource language pairs. As we will describe later in the paper, our comparable sentence classifier relies on the availability of an ini-

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 69–77,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

tial seed lexicon that can either be provided by a human or can be statistically trained from parallel corpora (Och and Ney, 2003). Experiments show that a broad coverage lexicon provides us with better coverage for effective identification of comparable corpora. However, availability of such a resource can not be expected in very low-resource language pairs, or even if present may not be of good quality. This opens an interesting research question - Can we also elicit such information effectively at low costs? We propose active learning strategies for identifying the most informative comparable sentence pairs which a human can then extract parallel segments from.

While the first form of supervision provides us with class labels that can be used for tuning the feature weights of our classifier, the second form of supervision enables us to better estimate the feature functions. For the comparable sentence classifier to perform well, we show that both forms of supervision are needed and we introduce an active learning protocol to combine the two forms of supervision under a single joint active learning strategy.

The rest of the paper is organized as follows. In Section 2 we survey earlier research as relevant to the scope of the paper. In Section 3 we discuss the supervised training setup for our classifier. In Section 4 we discuss the application of active learning to the classification task. Section 5 discusses the case of active learning with two different annotations and proposes an approach for combining them. Section 6 presents experimental results and the effectiveness of the active learning strategies. We conclude with further discussion and future work.

2 Related Work

There has been a lot of interest in using comparable corpora for MT, primarily on extracting parallel sentence pairs from comparable sources (Zhao and Vogel, 2002; Fung and Yee, 1998). Some work has gone beyond this focussing on extracting subsentential fragments from noisier comparable data (Munteanu and Marcu, 2006; Quirk et al., 2007). The research conducted in this paper has two primary contributions and so we will discuss the related work as relevant to each of them.

Our first contribution in this paper is the application of active learning for acquiring comparable data in the low-resource scenario, especially relevant when working with low-resource languages. There is some earlier work highlighting the need for techniques to deal with low-resource scenarios.(Munteanu and Marcu, 2005) propose bootstrapping using an existing classifier for collecting new data. However, this approach works when there is a classifier of reasonable performance. In the absence of parallel corpora to train lexicons human constructed dictionaries were used as an alternative which may, however, not be available for a large number of languages. Our proposal of active learning in this paper is suitable for highly impoverished scenarios that require support from a human.

The second contribution of the paper is to extend the traditional active learning setup that is suitable for eliciting a single annotation. We highlight the needs of the comparable corpora scenario where we have two kinds of annotations - class label assignment and parallel segment extraction and propose strategies in active learning that involve multiple annotations. A relevant setup is *multitask learning* (Caruana, 1997) which is increasingly becoming popular in natural language processing for learning from multiple learning tasks. There has been very less work in the area of multitask active learning. (Reichart et al., 2008) 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. (Settles et al., 2008) propose elicitation of annotations for image segmentation under a multi-instance learning framework.

Active learning with multiple annotations also has similarities to the recent body of work in learning from instance feedback and feature feedback (Melville et al., 2005). (Druck et al., 2009) propose active learning extensions to the gradient approach of learning from feature and instance feedback. However, in the comparable corpora problem although the second annotation is geared towards learning better features by enhancing the coverage of the lexicon, the annotation itself is not on the features but for extracting training data that is then used to train the lexicon.

3 Supervised Comparable Sentence Classification

In this section we discuss our supervised training setup and the classification algorithm. Our classifier tries to identify comparable sentences from among a large pool of noisy comparable sentences. In this paper we define comparable sentences as being translations that have around fifty percent or more translation equivalence. In future we will evaluate the robustness of the classifier by varying levels of noise at the sentence level.

3.1 Training the Classifier

Following (Munteanu and Marcu, 2005), we use a Maximum Entropy classifier to identify comparable sentences. The classifier probability can be defined as:

$$Pr(c_i|S,T) = \frac{1}{Z(S,T)} exp\left(\sum_{j=1}^{n} \lambda_j f_{ij}(c_i,S,T)\right)$$

where (S,T) is a sentence pair, c_i is the class, f_{ij} are feature functions and $Z(S)$ is a normalizing factor. The parameters λ_i are the weights for the feature functions and are estimated by optimizing on a training data set. For the task of classifying a sentence pair, there are two classes, $c_0 = comparable$ and $c_1 = non\ parallel$. A value closer to one for $Pr(c_1|S,T)$ indicates that (S,T) are comparable.

To train the classifier we need comparable sentence pairs and non-parallel sentence pairs. While it is easy to find negative examples online, acquiring comparable sentences is non-trivial and requires human intervention. (Munteanu and Marcu, 2005) construct negative examples automatically from positive examples by pairing all source sentences with all target sentences. We, however, assume the availability of both positive and negative examples to train the classifier. We use the GIS learning algorithm for tuning the model parameters.

3.2 Feature Computation

The features are defined primarily based on translation lexicon probabilities. Rather than computing word alignment between the two sentences, we use lexical probabilities to determine alignment points as follows: a source word s is aligned to a target word t if $p(s|t) > 0.5$. Target word alignment is computed similarly. Long contiguous sections of aligned words indicate parallelism. We use the following features:

- Source and target sentence length ratio
- Source and target sentence length difference
- Lexical probability score, similar to IBM model 1
- Number of aligned words
- Longest aligned word sequence
- Number of un-aligned words

Lexical probability score, and alignment features generate two sets of features based on translation lexica obtained by training in both directions. Features are normalized with respect to the sentence length.

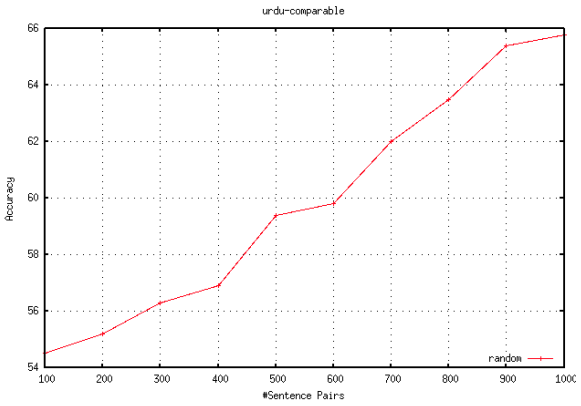

Figure 1: Seed parallel corpora size vs. Classifier performance in Urdu-English language pair

In our experiments we observe that the most informative features are the ones involving the probabilistic lexicon. However, the comparable corpora obtained for training the classifier cannot be used for automatically training a lexicon. We, therefore, require the availability of an initial seed parallel corpus that can be used for computing the lexicon and the associated feature functions. We notice that the size of the seed corpus has a large influence on the accuracy of the classifier. Figure 1 shows a plot with

the initial size of the corpus used to construct the probabilistic lexicon on x-axis and its effect on the accuracy of the classifier on y-axis. The sentences were drawn randomly from a large pool of Urdu-English parallel corpus and it is clear that a larger pool of parallel sentences leads to a better lexicon and an improved classifier.

4 Active Learning with Multiple Annotations

4.1 Cost Motivation

Lack of existing annotated data requires reliable human annotation that is expensive and effort-intensive. We propose active learning for the problem of effectively acquiring multiple annotations starting with unlabeled data. 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 a human expert, which it then uses to retrain the underlying supervised model for improving performance.

A meaningful setup to study multi annotation active learning is to take into account the cost involved for each of the annotations. In the case of comparable corpora we have two annotation tasks, each with cost models $Cost_1$ and $Cost_2$ respectively. The goal of multi annotation active learning is to select the optimal set of instances for each annotation so as to maximize the benefit to the classifier. Unlike the traditional active learning, where we optimize the number of instances we label, here we optimize the selection under a provided budget B_k per iteration of the active learning algorithm.

4.2 Active Learning Setup

We now discuss our active learning framework for building comparable corpora as shown in Algorithm 1. We start with an unlabeled dataset $U_0 = \{x_j =< s_j, t_j >\}$ and a seed labeled dataset $L_0 = \{(< s_j, t_j >, c_i)\}$, where $c \in 0, 1$ are class labels with 0 being the non-parallel class and 1 being the comparable data class. We also have $T_0 = \{< s_k, t_k >\}$ which corresponds to parallel segments or sentences identified from L_0 that will be used in training the probabilistic lexicon. Both T_0 and L_0 can be very small in size at the start of the active learning loop. In our experiments, we tried with as few as 50 to 100 sentences for each of the datasets.

We perform an iterative budget motivated active learning loop for acquiring labeled data over k iterations. We start the active learning loop by first training a lexicon with the available T_k and then using that we train the classifier over L_k. We, then score all the sentences in the U_k using the model θ and apply our selection strategy to retrieve the best scoring instance or a small batch of instances. In the simplest case we annotate this instance and add it back to the tuning set C_k for re-training the classifier. If the instance was a comparable sentence pair, then we could also perform the second annotation conditioned upon the availability of the budget. The identified sub-segments (ss_i , tt_i) are added back to the training data T_k used for training the lexicon in the subsequent iterations.

Algorithm 1 ACTIVE LEARNING SETUP

```
Given Unlabeled Comparable Corpus: U_0
Given Seed Parallel Corpus: T_0
Given Tuning Corpus: L_0
for k = 0 to K do
    Train Lexicon using T_k
    θ = Tune Classifier using C_k
    while Cost < B_k do
        i = Query(U_k,L_k,T_k,θ)
        c_i = Human Annotation-1 (s_i, t_i)
        (ss_i ,tt_i) = Human Annotation-2 x_i
        L_k = C_k ∪ (s_i, t_i, c_i)
        T_k = T_k ∪ (ss_i, tt_i)
        U_k = U_k - x_i
        Cost = Cost_1 + Cost_2
    end while
end for
```

5 Sampling Strategies for Active Learning

5.1 Acquiring Training Data for Classifier

Our selection strategies for obtaining class labels for training the classifier uses the model in its current state to decide on the informative instances for the next round of iterative training. We propose the following two sampling strategies for this task.

5.1.1 Certainty Sampling

This strategy selects instances where the current model is highly confident. While this may seem redundant at the outset, we argue that this criteria can be a good sampling strategy when the classifier is weak or trained in an impoverished data scenario. Certainty sampling strategy is a lot similar to the idea of unsupervised approaches like boosting or self-training. However, we make it a semi-supervised approach by having a human in the loop to provide affirmation for the selected instance. Consider the following scenario. If we select an instance that our current model prefers and obtain a contradicting label from the human, then this instance has a maximal impact on the decision boundary of the classifier. On the other hand, if the label is reaffirmed by a human, the overall variance reduces and in the process, it also helps in assigning higher preference for the configuration of the decision boundary. (Melville et al., 2005) introduce a certainty sampling strategy for the task of feature labeling in a text categorization task. Inspired by the same we borrow the name and also apply this as an instance sampling approach. Given an instance x and the classifier posterior distribution for the classes as $P(.)$, we select the most informative instance as follows:

$$x* = arg\,max_x P(c = 1|x)$$

5.1.2 Margin-based Sampling

The certainty sampling strategy only considers the instance that has the best score for the comparable sentence class. However we could benefit from information about the second best class assigned to the same instance. In the typical multi-class classification problems, earlier work shows success using such a 'margin based' approach (Scheffer et al., 2001), where the difference between the probabilities assigned by the underlying model to the first best and second best classes is used as the sampling criteria.

Given a classifier with posterior distribution over classes for an instance $P(c = 1|x)$, the margin based strategy is framed as $x* = arg\,min_x P(c_1|x) - P(c_2|x)$, where c_1 is the best prediction for the class and c_2 is the second best prediction under the model. It should be noted that for binary classification tasks with two classes, the margin sampling approach reduces to an uncertainty sampling approach (Lewis and Catlett, 1994).

5.2 Acquiring Parallel Segments for Lexicon Training

We now propose two sampling strategies for the second annotation. Our goal is to select instances that could potentially provide parallel segments for improved lexical coverage and feature computation.

5.2.1 Diversity Sampling

We are interested in acquiring clean parallel segments for training a lexicon that can be used in feature computation. It is not clear how one could use a comparable sentence pair to decide the potential for extracting a parallel segment. However, it is highly likely that if such a sentence pair has new coverage on the source side, then it increases the chances of obtaining new coverage. We, therefore, propose a diversity based sampling for extracting instances that provide new vocabulary coverage . The scoring function $tc\_score(s)$ is defined below, where $Voc(s)$ is defined as the vocabulary of source sentence s for an instance $x_i =< s_i, t_i >$, T is the set of parallel sentences or segments extracted so far.

$$tc\_score(s) = \sum_{s=1}^{|T|} sim(s, s') * \frac{1}{|T|} \quad (1)$$

$$sim(s, s') = |(Voc(s) \cap Voc(s')| \quad (2)$$

5.2.2 Alignment Ratio

We also propose a strategy that provides direct insight into the coverage of the underlying lexicon and prefers a sentence pair that is more likely to be comparable. We call this *alignment ratio* and it can be easily computed from the available set of features discussed in Section 3 as below:

$$a\_score(s) = \frac{\#unalignedwords}{\#alignedwords} \quad (3)$$

$$s* = arg\,max_s a\_score(s) \quad (4)$$

This strategy is quite similar to the diversity based approach as both prefer selecting sentences that have

a potential to offer new vocabulary from the comparable sentence pair. However while the diversity approach looks only at the source side coverage and does not depend upon the underlying lexicon, the alignment ratio utilizes the model for computing coverage. It should also be noted that while we have coverage for a word in the sentence pair, it may not make it to the probabilistically trained and extracted lexicon.

5.3 Combining Multiple Annotations

Finally, given two annotations and corresponding sampling strategies, we try to jointly select the sentence that is best suitable for obtaining both the annotations and is maximally beneficial to the classifier. We select a single instance by combining the scores from the different selection strategies as a geometric mean. For instance, we consider a margin based sampling ($margin$) for the first annotation and a diversity sampling (tc_score) for the second annotation, we can jointly select a sentence that maximizes the combined score as shown below:

$$total\_score(s) = margin(s) * tc\_score(s) \quad (5)$$

$$s* = arg\,max_s total\_score(s) \quad (6)$$

6 Experiments and Results

6.1 Data

This research primarily focuses on identifying comparable sentences from a pool of dominantly non-parallel sentences. To our knowledge, there is a dearth of publicly available comparable corpora of this nature. We, therefore, simulate a low-resource scenario by using realistic assumptions of noise and parallelism at both the corpus-level and the sentence-level. In this section we discuss the process and assumptions involved in the creation of our datasets and try to mimic the properties of real-world comparable corpora harvested from the web.

We first start with a sentence-aligned parallel corpus available for the language pair. We then divide the corpus into three parts. The first part is called the 'sampling pool' and is set aside to use for drawing sentences at random. The second part is used to act as a non-parallel corpus. We achieve non-parallelism by randomizing the mapping of the target sentences with the source sentences. This is a slight variation of the strategy used in (Munteanu and Marcu, 2005) for generating negative examples for their classifier. The third part is used to synthesize a comparable corpus at the sentence-level. We perform this by first selecting a parallel sentence-pair and then padding either sides by a source and target segment drawn independently from the sampling pool. We control the length of the non-parallel portion that is appended to be lesser than or equal to the original length of the sentence. Therefore, the resulting synthesized comparable sentence pairs are guaranteed to contain at least 50% parallelism.

We use this dataset as the unlabeled pool from which the active learner selects instances for labeling. Since the gold-standard labels for this corpus are already available, which gives us better control over automating the active learning process, which typically requires a human in the loop. However, our active learning strategies are in no way limited by the simulated data setup and can generalize to the real world scenario with an expert providing the labels for each instance.

We perform our experiments with data from two language pairs: Urdu-English and Spanish-English. For Urdu-English, we use the parallel corpus NIST 2008 dataset released for the translation shared task. We start with 50,000 parallel sentence corpus from the released training data to create a corpus of 25,000 sentence pairs with 12,500 each of comparable and non-parallel sentence pairs. Similarly, we use 50,000 parallel sentences from the training data released by the WMT 2008 datasets for Spanish-English to create a corpus of 25,000 sentence pairs. We also use two held-out data sets for training and tuning the classifier, consisting of 1000 sentence pairs (500 non-parallel and 500 comparable).

6.2 Results

We perform two kinds of evaluations: the first, to show that our active learning strategies perform well across language pairs and the second, to show that multi annotation active learning leads to a good improvement in performance of the classifier.

6.2.1 How does the Active Learning perform?

In section 5, we proposed multiple active learning strategies for both eliciting both kinds of annotations. A good active learning strategy should select

instances that contribute to the maximal improvement of the classifier. The effectiveness of active learning is typically tested by the number of queries the learner asks and the resultant improvement in the performance of the classifier. The classifier performance in the comparable sentence classification task can be computed as the F-score on the held out dataset. For this work, we assume that both the annotations require the same effort level and so assign uniform cost for eliciting each of them. Therefore the number of queries is equivalent to the total cost of supervision.

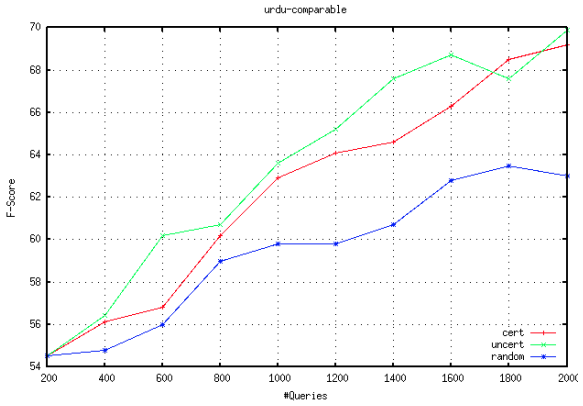

Figure 2: Active learning performance for the comparable corpora classification in Urdu-English language-pair

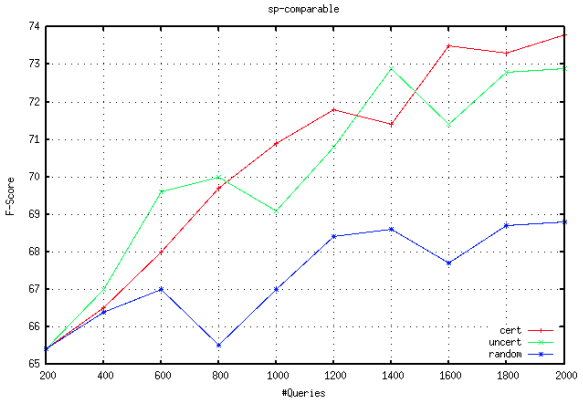

Figure 3: Active learning performance for the comparable corpora classification in Spanish-English language-pair

Figure 2 shows our results for the Urdu-English language pair, and Figure 3 plots the Spanish-English results with the x-axis showing the total number of queries posed to obtain annotations and the y-axis shows the resultant improvement in accuracy of the classifier. In these experiments we do not actively select for the second annotation but acquire the parallel segment from the same sentence. We compare this over a random baseline where the sentence pair is selected at random and used for eliciting both annotations at the same time.

Firstly, we notice that both our active learning strategies: certainty sampling and margin-based sampling perform better than the random baseline. For the Urdu-English language pair we can see that for the same effort expended (i.e 2000 queries) the classifier has an increase in accuracy of 8 absolute points. For Spanish-English language pair the accuracy improvement is 6 points over random baseline. Another observation from Figure 3 is that for the classifier to reach an fixed accuracy of 68 points, the random sampling method requires 2000 queries while the from the active selection strategies require significantly less effort of about 500 queries.

6.2.2 Performance of Joint Selection with Multiple Annotations

We now evaluate our joint selection strategy that tries to select the best possible instance for both the annotations. Figure 4 shows our results for the Urdu-English language pair, and Figure 5 plots the Spanish-English results for active learning with multiple annotations. As before, the x-axis shows the total number of queries posed, equivalent to the cumulative effort for obtaining the annotations and the y-axis shows the resultant improvement in accuracy of the classifier.

We evaluate the multi annotation active learning against two single-sided baselines where the sampling focus is on selecting instances according to strategies suitable for one annotation at a time. The best performing active learning strategy for the class label annotations is the certainty sampling (annot1) and so for one single-sided baseline, we use this baseline. We also obtain the second annotation for the same instance. By doing so, we might be selecting an instance that is sub-optimal for the second annotation and therefore the resultant lexicon may not maximally benefit from the instance. We also observe, from our experiments, that the diversity based sampling works well for the second anno-

tation and alignment ratio does not perform as well. So, for the second single-sided baseline we use the diversity based sampling strategy (annot2) and get the first annotation for the same instance. Finally we compare this with the joint selection approach proposed earlier that combines both the annotation strategies (annot1+annot2). In both the language pairs we notice that joint selection for both annotations performs better than the baselines.

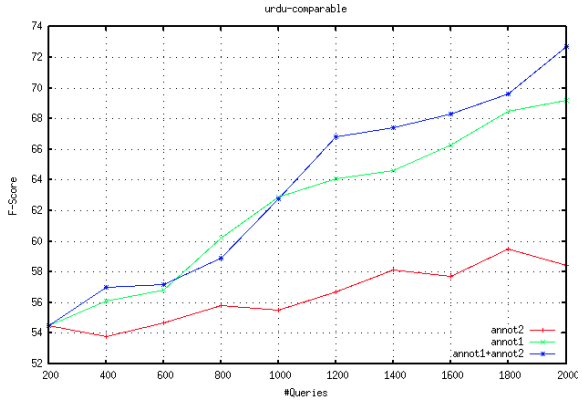

Figure 4: Active learning with multiple annotations and classification performance in Urdu-English

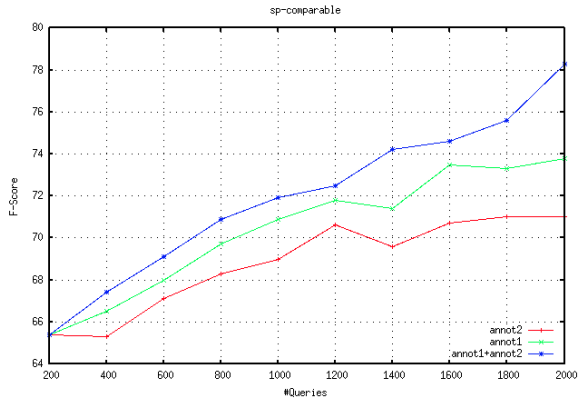

Figure 5: Active learning with multiple annotations and classification performance in Spanish-English

7 Conclusion and Future Work

In this paper, we proposed active learning with multiple annotations for the challenge of building comparable corpora in low-resource scenarios. In particular, we identified two kinds of annotations: class labels (for identifying comparable vs. non-parallel data) and clean parallel segments within the comparable sentences. We implemented multiple independent strategies for obtaining each of the abve in a cost-effective manner. Our active learning experiments in a simulated low-resource comparable corpora scenario across two language pairs show significant results over strong baselines. Finally we also proposed a joint selection strategy that selects a single instance which is beneficial to both the annotations. The results indicate an improvement over single strategy baselines.

There are several interesting questions for future work. Throughout the paper we assumed uniform costs for both the annotations, which will need to be verified with human subjects. We also hypothesize that obtaining both annotations for the same sentence may be cheaper than getting them from two different sentences due to the overhead of context switching. Another assumption is that of the existence of a single contiguous parallel segment in a comparable sentence pair, which needs to be verified for corpora on the web.

Finally, active learning assumes availability of an expert to answer the queries. Availability of an expert for low-resource languages and feasibility of running large scale experiments is difficult. We, therefore, have started working on crowdsourcing these annotation tasks on Amazon Mechanical Turk (MTurk) where it is easy to find people and quickly run experiments with real people.

Acknowledgement

This material is based upon work supported in part by the U. S. Army Research Laboratory and the U. S. Army Research Office under grant W911NF-10-1-0533, and in part by NSF under grant IIS 0916866.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Rich Caruana. 1997. Multitask learning. In *Machine Learning*, pages 41–75.

Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of Conference on Empirical Methods in Nat-*

ural Language Processing (EMNLP 2009), pages 81–90.

Jenny Rose Finkel and Christopher D. Manning. 2010. Hierarchical joint learning: Improving joint parsing and named entity recognition with non-jointly labeled data. In *Proceedings of ACL 2010*.

Pascale Fung and Lo Yen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 414–420, Montreal, Canada.

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

Prem Melville, Foster Provost, Maytal Saar-Tsechansky, and Raymond Mooney. 2005. Economical active feature-value acquisition through expected utility estimation. In *UBDM '05: Proceedings of the 1st international workshop on Utility-based data mining*, pages 10–16, New York, NY, USA. ACM.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 81–88, Sydney, Australia.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October.

Chris Quirk, Raghavendra U. Udupa, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of the Machine Translation Summit XI*, pages 377–384, Copenhagen, Denmark.

Roi Reichart, Katrin Tomanek, Udo Hahn, and Ari Rappoport. 2008. Multi-task active learning for linguistic annotations. In *Proceedings of ACL-08: HLT*, pages 861–869, Columbus, Ohio, June. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith. 2003. The web as a parallel corpus. *Comput. Linguist.*, 29(3):349–380.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *IDA '01: Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, pages 309–318, London, UK. Springer-Verlag.

Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-instance active learning. In *In Advances in Neural Information Processing Systems (NIPS*, pages 1289–1296. MIT Press.

Bing Zhao and Stephan Vogel. 2002. Full-text story alignment models for chinese-english bilingual news corpora. In *Proceedings of the ICSLP '02*, September.