Active Learning-Based Elicitation for Semi-Supervised Word Alignment

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

Abstract

Semi-supervised word alignment aims to improve the accuracy of automatic word alignment by incorporating full or partial manual alignments. Motivated by standard active learning query sampling frameworks like uncertainty-, margin- and query-by-committee sampling we propose multiple query strategies for the alignment link selection task. Our experiments show that by active selection of uncertain and informative links, we reduce the overall manual effort involved in elicitation of alignment link data for training a semi-supervised word aligner.

1 Introduction

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. 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 typically quite low 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, 2007a). 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.

Active learning for MT has not yet been explored to its full potential. Much of the literature has explored one task – selecting sentences to translate and add to the training corpus (Haffari and Sarkar, 2009). In this paper we explore active learning for word alignment, where the input to the active learner is a sentence pair (S, T) and the annotation elicited from human is a set of links $\{a_{ij}, \forall s_i \in S, t_j \in T\}$. Unlike previous approaches, our work does not require elicitation of full alignment for the sentence pair, which could be effort-intensive. We propose active learning query strategies to selectively elicit partial alignment information. Experiments in Section 5 show that our selection strategies reduce alignment error rates significantly over baseline.

2 Related Work

Researchers have begun to explore models that use both labeled and unlabeled data to build word-alignment models for MT. Fraser and Marcu (2006) pose the problem of alignment as a search problem in log-linear space with features coming from the IBM alignment models. The log-

Proceedings of the ACL 2010 Conference Short Papers, pages 365–370,
Uppsala, Sweden, 11-16 July 2010.

linear model is trained on available labeled data to improve performance. They propose a semi-supervised training algorithm which alternates between discriminative error training on the labeled data to learn the weighting parameters and maximum-likelihood EM training on unlabeled data to estimate the parameters. Callison-Burch et al. (2004) also improve alignment by interpolating human alignments with automatic alignments. They observe that while working with such data sets, alignments of higher quality should be given a much higher weight than the lower-quality alignments. Wu et al. (2006) learn separate models from labeled and unlabeled data using the standard EM algorithm. The two models are then interpolated to use as a learner in the semi-supervised algorithm to improve word alignment. To our knowledge, there is no prior work that has looked at reducing human effort by selective elicitation of partial word alignment using active learning techniques.

3 Active Learning for Word Alignment

Active learning attempts to optimize performance by selecting the most informative instances to label where 'informativeness' is defined as maximal expected improvement in accuracy. The objective is to select optimal instance for an external expert to label and then run the learning method on the newly-labeled and previously-labeled instances to minimize prediction or translation error, repeating until either the maximal number of external queries is reached or a desired accuracy level is achieved. Several studies (Tong and Koller, 2002; Nguyen and Smeulders, 2004; Donmez and Carbonell, 2008) show that active learning greatly helps to reduce the labeling effort in various classification tasks.

3.1 Active Learning Setup

We discuss our active learning setup for word alignment in Algorithm 1. 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 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.

Algorithm 1 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 = 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**

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, 2007b) 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.

3.2 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 all positions 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.

4 Query Strategies for Link Selection

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 θ_t for scoring each link for uncertainty, which consists of bidirectional translation lexicon tables computed from the bidirectional alignments.

4.1 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:

$$Score(a_{ij}/s_1^I, t1^J) = \frac{2 * P(t_j/s_i) * P(s_i/t_j)}{P(t_j/s_i) + P(s_i/t_j)} \quad (1)$$

4.2 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_1^I, t_1^J) and its word alignment, we compute two confidence metrics at alignment link level – based on the posterior link probability as seen in Equation 5. 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.

$$P_{t2s}(a_{ij}, t_1^J/s_1^I) = \frac{p_{t2s}(t_j/s_i, a_{ij} \in A)}{\sum_i^M p_{t2s}(t_j/s_i)} \quad (2)$$

$$P_{s2t}(a_{ij}, s_1^I/t_1^J) = \frac{p_{s2t}(s_i/t_j, a_{ij} \in A)}{\sum_i^N p_{s2t}(s_i/t_j)} \quad (3)$$

$$Conf1(a_{ij}/S, T) = \frac{2 * P_{t2s} * P_{s2t}}{P_{t2s} + P_{s2t}} \quad (4)$$

$$(5)$$

4.3 Query by Committee

The generative alignments produced differ based on the choice of direction of the language pair. We use A_{s2t} to denote alignment in the source to target direction and A_{t2s} 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 6. 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.

$$Score(a_{ij}) = \alpha \quad (6)$$

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

4.4 Margin Sampling

The strategy for confidence based sampling only considers information about the best scoring link

$conf(a_{ij}/S,T)$. However we could benefit from information about the second best scoring link as well. In 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 labels is used as a sampling criteria. We adapt such a margin-based approach to link-selection using the $Conf1$ scoring function discussed in the earlier sub-section. Our $margin$ technique is formulated below, where $\hat{a1}_{ij}$ and $\hat{a2}_{ij}$ 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.

$$Margin(i) = Conf1(\hat{a1}_{ij}/S,T) \quad -Conf1(\hat{a2}_{ij}/S,T)$$

5 Experiments

5.1 Data Setup

Our aim in this paper is to show that active learning can help select the most informative alignment links that have high uncertainty according to a given automatically trained model. We also show that fixing such alignments leads to the maximum reduction of error in word alignment, as measured by AER. 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.

5.2 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 corpus.

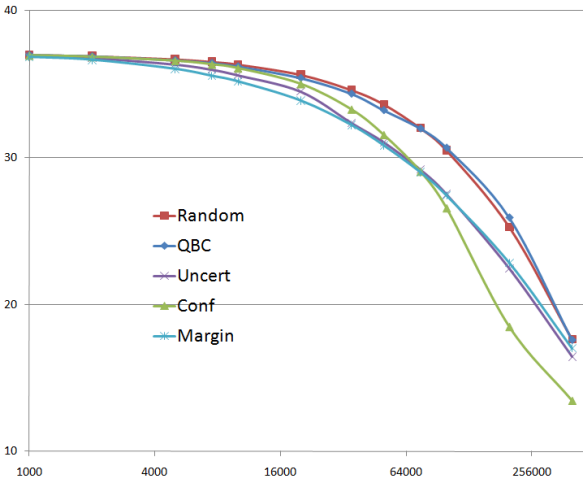

Figure 1: Performance of active sampling strategies for link selection

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 1 [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 MT-Eval 2004 dataset and test on a subset of MT-Eval 2004 dataset consisting of 631 sentences. We first obtain 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

[1] X axis has number of links elicited on a log-scale

System	BLEU	METEOR
Baseline	18.82	42.70
Human Alignment	19.96	44.22
Active Selection 20%	19.34	43.25

Table 1: Alignment and Translation Quality

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.

5.3 Batch Selection

Re-training the word alignment models after eliciting every individual alignment link is infeasible. In our data set of 21,863 sentences with 588,075 links, it would be computationally intensive to re-train after eliciting even 100 links in a batch. We therefore sample links as a discrete batch, and train alignment models to report performance at fixed points. Such a batch selection is only going to be sub-optimal as the underlying model changes with every alignment link and therefore becomes 'stale' for future selections. We observe that in some scenarios while fixing one alignment link could potentially fix all the mis-alignments in a sentence pair, our batch selection mechanism still samples from the rest of the links in the sentence pair. We experimented with an exponential decay function over the number of links previously selected, in order to discourage repeated sampling from the same sentence pair. We performed an experiment by selecting one of our best performing selection strategies ($conf$) and ran it in both configurations - one with the decay parameter ($batchdecay$) and one without it ($batch$). As seen in Figure 2, the decay function has an effect in the initial part of the curve where sampling is sparse but the effect gradually fades away as we observe more samples. In the reported results we do not use batch decay, but an optimal estimation of 'staleness' could lead to better gains in batch link selection using active learning.

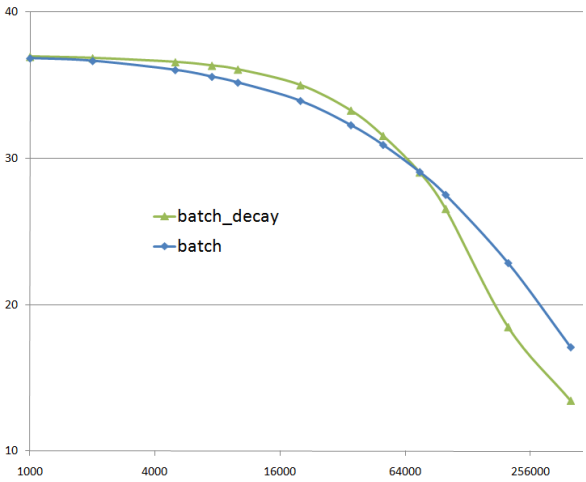

Figure 2: Batch decay effects on Conf-posterior sampling strategy

6 Conclusion and Future Work

Word-Alignment is a particularly challenging problem and has been addressed in a completely unsupervised manner thus far (Brown et al., 1993). While generative alignment models have been successful, lack of sufficient data, model assumptions and local optimum during training are well known problems. Semi-supervised techniques use partial manual alignment data to address some of these issues. We have shown that active learning strategies can reduce the effort involved in eliciting human alignment data. The reduction in effort is due to careful selection of maximally uncertain links that provide the most benefit to the alignment model when used in a semi-supervised training fashion. Experiments on Chinese-English have shown considerable improvements. In future we wish to work with word alignments for other language pairs like Arabic and English. We have tested out the feasibility of obtaining human word alignment data using Amazon Mechanical Turk and plan to obtain more data reduce the cost of annotation.

Acknowledgments

This research was partially supported by DARPA under grant NBCHC080097. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the DARPA. The first author would like to thank Qin Gao for the semi-supervised word alignment software and help with running experiments.

References

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2004. Confidence estimation for machine translation. In *Proceedings of Coling 2004*, pages 315–321, Geneva, Switzerland, Aug 23–Aug 27. COLING.

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

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In *ACL 2004*, page 175, Morristown, NJ, USA. Association for Computational Linguistics.

Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 105–112, Morristown, NJ, USA.

Pinar Donmez and Jaime G. Carbonell. 2008. Optimizing estimated loss reduction for active sampling in rank learning. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 248–255, New York, NY, USA. ACM.

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 769–776, Morristown, NJ, USA. Association for Computational Linguistics.

Alexander Fraser and Daniel Marcu. 2007a. Getting the structure right for word alignment: LEAF. In *Proceedings of the 2007 Joint Conference on EMNLP-CoNLL*, pages 51–60.

Alexander Fraser and Daniel Marcu. 2007b. Measuring word alignment quality for statistical machine translation. *Comput. Linguist.*, 33(3):293–303.

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

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, June. Association for Computational Linguistics.

Gholamreza Haffari and Anoop Sarkar. 2009. Active learning for multilingual statistical machine translation. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 181–189, Suntec, Singapore, August. Association for Computational Linguistics.

Fei Huang. 2009. Confidence measure for word alignment. In *Proceedings of the Joint ACL and IJCNLP*, pages 932–940, Suntec, Singapore, August. Association for Computational Linguistics.

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

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

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT 2007*, pages 228–231, Morristown, NJ, USA.

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.

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

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, pages 19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL 2002*, pages 311–318, Morristown, NJ, USA.

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.

Simon Tong and Daphne Koller. 2002. Support vector machine active learning with applications to text classification. *Journal of Machine Learning*, pages 45–66.

Nicola Ueffing and Hermann Ney. 2007. Word-level confidence estimation for machine translation. *Comput. Linguist.*, 33(1):9–40.

Hua Wu, Haifeng Wang, and Zhanyi Liu. 2006. Boosting statistical word alignment using labeled and unlabeled data. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 913–920, Morristown, NJ, USA. Association for Computational Linguistics.