Semi-Supervised Block ITG Models for Word Alignment

Gholamreza Haffari **Majid Razmara** **Fred Popowich**
School of Computing Sciences
Simon Fraser University
{ghaffar1,mra44,popowich}@cs.sfu.ca

1 Introduction

Labeled training data for the word alignment task, in the form of word-aligned sentence pairs, is hard to come by for many language-pairs. Hence, it is natural to draw upon semi-supervised learning methods (Fraser and Marcu, 2006). We introduce a semi-supervised learning method for word alignment using conditional entropy regularization (Grandvalet and Bengio, 2005) on top of a BITG-based discriminative model. Our preliminary experiments show improvement in the alignment quality compared to a strong supervised model (Haghighi et al., 2009).

Let $\mathcal{L} = \{\langle \mathbf{x}_i, \mathbf{y}_i \rangle\}_1^L$ be a set of labeled examples where $\mathbf{x}_i$ is an input and $\mathbf{y}_i$ is its output label, and $\mathcal{U} = \{\mathbf{x}_j\}_1^U$ be a set of unlabeled examples. The goal of semi-supervised learning is to take into account both labeled and unlabeled data in finding a good mapping from input to output. In the word alignment problem, the label $\mathbf{y}$ is the word alignment for the sentence pair in $\mathbf{x}$.

2 Word Alignment with Block ITG

Inversion transduction grammar (ITG) is a special synchronous context free grammar in which derivations of sentence pairs correspond to alignments (Wu, 1997). In its original form, there is only one nonterminal X and three possible rule types: (i) Terminal unary productions $X \rightarrow e/f$ where e and f are aligned source and target language word pair, (ii) Straight binary rule $X \rightarrow [X_1 X_2]$ where an aligned span is constructed from children as $X_1 X_2 / X_1 X_2$, (iii) Inverted binary rules $X \rightarrow \langle X_1 X_2 \rangle$ where the order of the children nonterminals are inverted $X_1 X_2 / X_2 X_1$. In block ITG (BITG), we allow the unary production rules to go to phrase pairs $X \rightarrow \bar{e}/\bar{f}$ instead of word pairs; note that the empty string (or null) is also considered a phrase.

BITG puts a structure over the alignment space and thus reorderings. Of course, this restricted space does not include all possible alignments, but it is shown that most of the alignments for some language pairs, such as French-English, can be covered by the BITG alignments (Cherry and Lin, 2006; L. Huang and Knight, 2009).

As in (Haghighi et al., 2009), we assume that the feature representation of an alignment $\phi(\mathbf{x}, \mathbf{y})$ decomposes over individual alignment links:

$$\phi(\mathbf{x}, \mathbf{y}) := \sum_{(\bar{e}, \bar{f}) \in A(\mathbf{x}, \mathbf{y})} \phi(\bar{e}, \bar{f})$$

where $\phi(\bar{e}, \bar{f})$ is the feature vector representation of a phrase pair, and $A(\mathbf{x}, \mathbf{y})$ is the set of phrase pairs produced for the sentence pair $\mathbf{x}$ according to the alignment $\mathbf{y}$. The score of an alignment for a sentence pair is the sum of scores for individual alignment links (or potentials), hence the best alignment $\mathbf{y}^*$ for a sentence pair (in the test time) is chosen by

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \boldsymbol{\theta} \cdot \phi(\mathbf{x}, \mathbf{y})$$

where $\boldsymbol{\theta}$ is the parameter vector. In what follows, we show how to learn the parameter vector $\boldsymbol{\theta}$ based on *both* labeled and unlabeled data.

3 Likelihood-based Semi-Supervised Training

We put a distribution over the alignments, and learn the parameters based on a likelihood objective function. We define a Gibbs distribution over the alignments in the space as $P_{\boldsymbol{\theta}}(\mathbf{y}|\mathbf{x}) := \frac{e^{\boldsymbol{\theta} \cdot \phi(\mathbf{x}, \mathbf{y})}}{Z_{\boldsymbol{\theta}}(\mathbf{x})}$ where $Z_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{\mathbf{y}} e^{\boldsymbol{\theta} \cdot \phi(\mathbf{x}, \mathbf{y})}$ is the so-called partition function to make the distribution sum to one.

One important idea in semi-supervised learning for probabilistic models is to prefer those parameter values which make the prediction of the model on the unlabeled data more *confident*. In a sense, unlabeled data is used to induce a *data-dependent* regularization on model parameters. Conditional entropy regularization is an instance of this methodology where the confidence is measured via an entropy measure. That is, the best value $\boldsymbol{\theta}^*$ for the parameter vector is found by

$$\arg\max_{\boldsymbol{\theta}} \frac{1}{L}\sum_{i=1}^{L}\log P_{\boldsymbol{\theta}}(\mathbf{y}_i|\mathbf{x}_i) - \frac{\gamma}{U}\sum_{j=1}^{U} R_\alpha(P(.|\mathbf{x}_j)) \quad (1)$$

where γ is a trade-off parameter, and R_α in our case is the family of Rényi entropy measures: $R_\alpha(P) = \frac{1}{1-\alpha}\log\left(\sum_{\mathbf{y}} P^\alpha(\mathbf{y})\right)$. It can be shown that $\lim_{\alpha\rightarrow 1} R_\alpha(P)$ corresponds to the Shannon entropy $-\sum_{\mathbf{y}} P(\mathbf{y})\log P(\mathbf{y})$, and $\lim_{\alpha\rightarrow\infty} R_\alpha(P)$ corresponds to $-\log\max_{\mathbf{y}} P(\mathbf{y})$, i.e. the negative log-probability of the modal or "Viterbi" label (Arndt, 2001). When $\alpha \rightarrow 0$, the Rényi entropy of a distribution approaches the (Shannon) entropy of the uniform distribution. Our use of Rényi entropy is inspired by (Smith and Eisner, 2007), who noticed that using Rényi entropy measures instead of the conventional Shannon entropy offers more flexibility in terms of the entropy measures while allows *efficient* algorithms for parameter estimation. We use gradient descent to optimize the training objective (1), and learn the parameters.

4 Experiments

In this section, we report some preliminary results on a subset of the English-French Hansards data set from the 2003 NAACL shared task [1]. The sizes of labeled/unlabeled/test datasets are 29/1533/252 sentence pairs, respectively. Evaluation is done based on the following measures: AER (alignment error rate), and F-score of the predicted alignment links. The supervised model (baseline) has 10.10% AER and 89.53% F-score on the test set. These performance measures are improved by our semi-supervised model to 9.45% AER and 90.25% F-score, where $\alpha = .01$ and $\gamma = 5$.

We also investigate the effect of different values for α in the Rényi entropy. Figure 1 shows the F-score of the semi-supervised trained model with different values for α with respect to γ (the x-axis) which controls the effect of unlabeled data in the training process. It is interesting to see that small values for α are less sensitive to variation in γ in producing good parameter values.

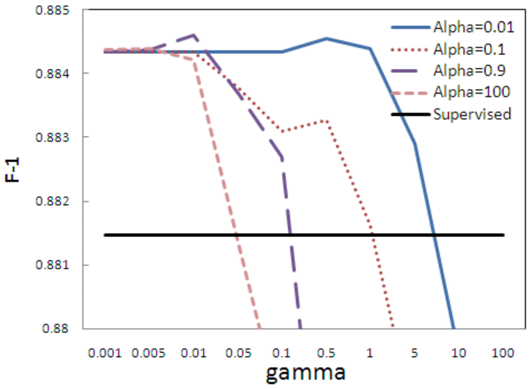

Figure 1: F-score of the semi-supervised trained model, evaluated on the labeled part of the training data.

5 Related Work

(Fraser and Marcu, 2006) experimented a semi-supervised learning on English-French and English-Arabic language pairs. They use a log-linear model which consisted of 5 sub-models of IBM Model 4 along with 11 other feature functions. First, sub-model parameters are initialized using the alignments generated by IBM Model 4, then using the MERT algorithm (Och, 2003), the sub-model contributions (feature weights) are estimated and iteratively a similar procedure to EM as well as MERT are applied to learn both sub-model parameters and contributions. Since enumerating all possible alignments is intractable, Viterbi EM training (approximate EM) have been used.

6 Future Work

We would like to apply our framework on large scale datasets and more language pairs, and augment our model with more rich features. The success of our approach will lead to building high quality word-alignment models for many language-pairs, and hopefully improving the translation quality.

References

C. Arndt. 2001. *Information Measures*. Springer.

[1] `http://www.cse.unt.edu/∼rada/wpt` .

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Comput. Linguist.*, 18(4):467–479.

C. Cherry and D. Lin. 2006. A comparison of syntactically motivated word alignment spaces. In *Proceedings of EACL*.

A. Fraser and D. Marcu. 2006. Semi-supervised training for statistical word alignment. In *Proceedings of ACL*.

Y. Grandvalet and Y. Bengio. 2005. Semi-supervised learning by entropy minimization. In *Proceedings of NIPS*.

A. Haghighi, J. Blitzer, J. DeNero, and D. Klein. 2009. Better word alignments with supervised itg models. In *Proceedings of ACL-AFNLP*.

D. Gildea L. Huang, H. Zhang and K. Knight. 2009. Binarization of synchronous context-free grammars. *Computational Linguistics, 35(4)*.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *ACL*, pages 160–167.

D. Smith and J. Eisner. 2007. Bootstrapping feature-rich dependency parsers with entropic priors. In *Proceedings of (EMNLP-CoNLL)*.

D. Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23.

A Cluster Based Features

We have used the Brown algorithm (Brown et al., 1992) for word clustering to add cluster-based features to our feature set. Using Percy Liang's C++ implementation of the algorithm[2] the words of the source and foreign languages are grouped into 50 classes each. Taking the top K frequent class pair in the unlabeled dataset as a new feature set, we run our code on the same dataset used before. However, it does not seem to be an improvement in the training or test sets.

The following table compares the mean and standard deviation of weights of lexical features with those of the new cluster-based features. The weights are taken from an experiment using the top 50 frequent cluster-based features and the same number of lexical features.

	Mean	Standard Deviation
Lexical Featuers	-0.0714	0.3858
Cluster-based Features	0.0657	0.1885

[2] Available on `http://www.eecs.berkeley.edu/~pliang/software/`