Discriminative Weighted Alignment Matrices For Statistical Machine Translation

Nadi Tomeh and **Alexandre Allauzen** and **François Yvon**
LIMSI-CNRS and Université Paris Sud
BP 133, 91403 Orsay
{nadi,allauzen,yvon}@limsi.fr

Abstract

In extant phrase-based statistical machine translation (SMT) systems, the translation model relies on word-to-word alignments, which serve as constraints for the subsequent heuristic extraction and scoring processes. Word alignments are usually inferred in a probabilistic framework; yet, only one single best alignment is retained, as if alignments were deterministically produced. In this paper, we explore ways to take into account the entire alignment matrix, where each alignment link is scored by its probability. By comparison with previous attempts, we use an exponential model to compute these probabilities, which enables us to achieve significant improvements on the *NIST MT'09* Arabic-English translation task.

1 Introduction

In Phrase-Based SMT systems, a source sentence is translated by concatenating translation options, selected from an inventory called the *phrase table*. Building this inventory from a parallel corpus constitute the *translation model training* phase, which is usually performed in two main steps. 1) For each training sentence pair, a set of source-target phrase-pairs, that are translations of one another, are first extracted. 2) Phrase pairs accumulated over the entire training corpus are collected and scored using relative frequencies estimates. The collection of phrase-pairs and their scores constitutes the translation model.

During the extraction step, we would like to use a phrase alignment model that enables the computation of corpus level statistics related to the joint segmentation and alignment of source and target sentences. Unfortunately, generative models designed for this purpose (Marcu and Wong, 2002; Birch et al., 2006) fail to deliver good performance due to three key difficulties (DeNero et al., 2006).

First, exploring the whole space of phrase-to-phrase alignment is intractable, which makes phrase alignment a NP-hard (DeNero and Klein, 2008) problem. Second, including a latent segmentation variable in the model increases the risk of overfitting during EM training. Third, spurious segmentation ambiguity tends to populate the phrase table with more entries, each having too few translation options. A practical solution is to reconfigure the phrase alignment problem in terms of *words* instead of phrases: a fixed segmentation, based on word boundaries, is used, and the resulting model is simpler to train using EM. Then, for each word-aligned sentence in the training corpus, an additional step is required to identify the set of phrase-pairs to be extracted. A heuristic which extracts phrase-pairs that are consistent with the Viterbi word alignment is widely used in practice.

During the scoring step, relative frequencies computed on the training corpus are used to assess each extracted phrase pair. Additional scores, based on lexical probabilities are also used so as to smooth the scores of rare phrase-pairs.

The training of the translation model is thus decomposed as a modular pipeline the components of which can be developed independently. The resulting modularity comes at the price of possible error propagation between consecutive steps: errors in the 1-best word alignment can propagate to phrase pair extraction and to probability estimation.

This problem can be alleviated by feeding more information from word alignment into the pipeline.

Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.)
Proceedings of the 15th Conference of the European Association for Machine Translation, p. 305–312
Leuven, Belgium, May 2011

For this purpose, a structure called the *Weighted Alignment Matrix* (WAM) (Liu et al., 2009), which compactly encodes the distribution of all possible alignments of a sentence pair, can be used to extract and score phrase-pairs. Each cell in this matrix corresponds to a pair of (source, target) words; the associated value measures the quality of the alignment link. Therefore, a weighted matrix encodes, in linear space, the probabilities of exponentially many alignments.

The authors of (Liu et al., 2009) estimate link probabilities by calculating relative frequencies over a list of N-best alignments produced by generative models, and show some improvements in translation quality. However, using small N-best lists as samples is known to yield poor estimates of the alignment posteriors, as these lists usually contain too few variations. In this paper, we argue that better estimation of alignment probabilities helps achieving clearer improvements. Our solution is to directly model the weighted alignment matrices using a discriminative aligner (Ayan and Dorr, 2006; Tomeh et al., 2010).

The rest of this paper is organized as follows: we start in Section 2 by a recap of related work. Section 3 revisits the standard translation model procedures and its extensions to weighted matrices. Our own approach is introduced in Section 4 and experimentally contrasted to various baselines in 5. We discuss further prospects in Section 6.

2 Related work

As pointed out in the introduction, the construction of the translation model starts with a word alignment step during which relevant phrase-pairs are extracted and their probabilities are estimated. Yet, word alignment outputs a probability distribution over all possible alignments. However, the most common practice (Koehn et al., 2003) is to use only the 1-best, Viterbi alignment, while discarding all the other informations contained in this distribution, which seems to adversely impact the quality of the resulting translation model.

In fact, several researchers have shown that incorporating more information from the posterior distribution helps reducing the propagation of errors and improves performance. In (Mi and Huang, 2008), some gains are achieved by exploiting a packed forest, which compactly encodes exponentially many parses, to extract rules for a syntax-based translation system, instead of using only the 1-best tree. This compact representation has already been shown to be efficient and effective (Galley et al., 2006; Wang et al., 2007).

Similarly, N-best alignments are used to extract phrase-pairs as in (Xue et al., 2006; Venugopal et al., 2008); in the latter, a probability distribution over N-best alignments and parses is used to generate posterior fractional counts for rules in a syntax-based translation model.

Due to the difficulty of computing statistics under IBM3 and IBM4 models, the previously described approaches use N-best alignments as samples to approximate word-to-word alignment posterior probabilities. While simpler models, such as HMM and IBM1, allow for such a computation (Brown et al., 1993; Venugopal et al., 2003; Deng and Byrne, 2005), they do not compete with Model 4 in terms of performance. A solution to this problem is described in (Deng and Byrne, 2005), where a *word-to-phrase* HMM alignment model is proposed, which constitutes a competitive model to IBM4. Under this model, the necessary statistics can be computed efficiently with the forward algorithm. The phrase pair induction procedure described in (Deng and Byrne, 2005), benefits from this efficiency to estimate a phrase-to-phrase posterior distribution, which is used further in the extraction and scoring of phrases. In (de Gispert et al., 2010), a similar procedure is shown to be useful for extracting synchronous grammar rules.

A structure analogous to the packed forest for trees is presented in (Liu et al., 2009) and called Weighted Alignment Matrix. Each element in the matrix is assigned a probability which measures the confidence of a word alignment. An algorithm for extracting phrase-pairs from weighted matrices and for estimating their scores is shown to be beneficial to translation quality.

In this paper, we continue this line of research and show that additional improvements can be obtained by better estimating the word alignments in a discriminative manner, using the MaxEnt-based word aligner described in (Ayan and Dorr, 2006; Tomeh et al., 2010; Tomeh et al., 2011).

3 WAM-based Translation Models

The translation model $\mathcal{T}$ constitutes the primary source of knowledge in a phrase-based SMT system and plays a crucial role in determining the quality of its output. Recall that a translation model is simply a list of bilingual phrase-pairs that

are translations of one another. Each phrase-pair is associated with a set of scores assessing its relevance for the translation task, where each score is based on statistics accumulated over some training corpus. In this section, we present a general framework which enables to frame both the standard and the WAM-based approaches.

3.1 A General Framework

Algorithm 1 sketches a general approach to construct the translation model $\mathcal{T}$, by extracting and scoring phrase-pairs from a parallel corpus $\mathcal{C}$.

For all sentence pairs (e_1^I, f_1^J) made up of J source words and I target words, we would like to enumerate all possible phrase-pairs $(f_{j_1}^{j_2}, e_{i_1}^{i_2})$ and assign each of them a score (f_E) that can be used to inform the selection criteria and/or provide a fractional count quantifying its quality (step 5).

Yet, extracting *all* possible phrase-pairs found in the training corpus would cause practical problems, as (1) the correlated growth in the solution space would dramatically slow down decoding; (2) the simplicity of the scoring procedure, based on relative frequencies, can not distinguish relevant rare translation candidates from noisy ones and would cause some probability mass to be wasted on noisy phrase-pairs. Hence the need for a selection procedure implemented in steps 4 and 6, which discard all phrase-pairs that do not satisfy some alignment constraints $\mathcal{C}_\mathcal{A}$ (based on the matrix $\mathbf{A}$) or do not fit some selection criteria $\mathcal{C}_\mathcal{S}$.

The final step is to add a set of scores to each selected phrase-pair (step 10). These scores usually include a translation probability ϕ, estimated using relative frequencies over the training corpus, where each occurence of a phrase-pair is evaluated using the counting function f_C. They also include lexical weights lex, based on lexical translation probabilities w, as a smoothing method to improve the estimates computed for rare phrase-pairs. A valuable, and relatively easy to acquire, source of information is the word alignment represented by the alignement matrix $\mathbf{A}$, which is consulted by the different steps of this algorithm: filtering, evaluation and scoring of phrase-pairs.

3.2 Standard Instantiation

The most common instantiation of this framework (Koehn et al., 2003) considers a binary alignmnent matrix $\mathbf{A}$, where each cell represents a binary variable indicating whether the associated words are aligned or not. The matrix is usually obtained

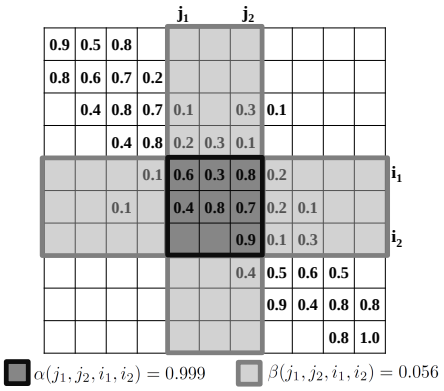

Figure 1: Computation of fractional counts: $f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2}) = \alpha(j_1, j_2, i_1, i_2) \times \beta(j_1, j_2, i_1, i_2)$. Empty cells have zero probability.

by applying the symmetrization heuristic to two Viterbi alignments, one for each translation direction. The alignment constraints $\mathcal{C}_\mathcal{A}$ are defined so that extracted phrase-pairs $(f_{j_1}^{j_2}, e_{i_1}^{i_2})$ are consistent with $\mathbf{A}$: $\forall(i,j) \in \mathbf{L} : (j \in [j_1, j_2] \wedge i \in [i_1, i_2]) \vee (j \notin [j_1, j_2] \wedge i \notin [i_1, i_2])$ where $\mathbf{L}$ is the set of all active links in $\mathbf{A}$. The selection criteria $\mathcal{C}_\mathcal{S}$ helps improve the practical efficiency by establishing a limit on the admissible source/target phrase lengths. All selected phrase-pairs are uniformly evaluated and counted using $f_E = f_C = 1$.

3.3 WAM-based Instantiation

Since the standard instantiation ignores alignment probabilities, it tends to be sensitive to alignment's precision and recall errors. An erroneous link, as unlikely as it may be, can prevent the extraction of many plausible phrase-pairs. Furthermore, the extracted phrase-pairs are all considered of equal quality, regardless of how much evidence the alignment matrix provides for them. A more flexible and robust alternative instantiation takes advantage of a structure called Weighted Alignment Matrix (WAM), presented in (Liu et al., 2009). In the weighted matrix $\mathbf{A_w} = \{p(a_{i,j}|\mathbf{e}, \mathbf{f}) : 1 \leq i \leq I, 1 \leq j \leq J\}$, each possible link is weighted by a score $p(a_{i,j}|\mathbf{e}, \mathbf{f})$ quantifying the confidence assigned to it by the alignment model.

Evaluation and Counting Functions The use of a weighted matrix allows for conceptualizing more informative evaluation and counting functions, which can help mitigate the error propagation problem. To incorporate alignment posterior probabilities when computing fractional counts for a phrase-pair, all possible alignments should be

Algorithm 1 Translation Model Construction

Input: Parallel Corpus $\mathcal{C}$

Output: Translation Model $\mathcal{T}$

1: Initialize the phrase table $\mathcal{P} = \{\}$

2: **for all** sentence pairs in the training parallel corpus $(e_1^I, f_1^J) \in \mathcal{C}$ **do**

3: Construct the alignment matrix $\mathbf{A} = align(e_1^I, f_1^J)$

4: $\mathcal{P}_\mathcal{A} = \left\{(f_{j_1}^{j_2}, e_{i_1}^{i_2}) : 1 \leq j \leq J, 1 \leq i \leq I, (f_{j_1}^{j_2}, e_{i_1}^{i_2}) \text{ satisfies some alignment constraints } \mathcal{C}_\mathcal{A}\right\}$

5: $\mathcal{P}_\mathcal{E} = \{\langle x, f_E(x, \mathbf{A})\rangle : x \in \mathcal{P}_\mathcal{A}\}$, where f_E is an evaluation function

6: $\mathcal{P}_\mathcal{S} = \{x : x \in \mathcal{P}_\mathcal{E}, x \text{ satisfies some selection criteria } \mathcal{C}_\mathcal{S}\}$

7: $\mathcal{P} = \mathcal{P} \cup \mathcal{P}_\mathcal{S}$

8: **end for**

9: **for all** $\langle (e, f), f_E(e, f)\rangle \in \mathcal{P}$ **do**

10: $\mathcal{T} = \mathcal{T} \cup \{\langle (e, f), \phi(e|f), \phi(f|e), lex(e|f), lex(f|e)\rangle\}$ where f_C is a counting function,

$$\phi(e|f) = \frac{f_C(e, f)}{\sum_{e_i} f_C(e_i, f)}, \text{ and } lex(e|f, a_{e,f}) = \prod_{i=1}^{length(e)} \frac{1}{|\{j : (i, j) \in a_{e,f}\}|} \sum_{\forall (i,j) \in a_{e,f}} w(e_i|f_j),$$

11: **end for**

explicitly enumerated. Unlike for N-best (Venugopal et al., 2008) or HMM (de Gispert et al., 2010) alignments, this is unrealistic for a weighted matrix. Instead, we follow (Liu et al., 2009) and use link probabilities to compute a fractional count, interpreted as the probability that the phrase-pair satisfies consistency constraints.

Given a weighted alignment matrix $\mathbf{A_w}$ and a phrase-pair $(f_{j_1}^{j_2}, e_{i_1}^{i_2})$, two regions (in gray on Figure 1) are identified: $in(j_1, j_2, i_1, i_2)$ and $out(j_1, j_2, i_1, i_2)$ which respectively represents links *inside* and *outside* (on the same rows and columns) of a phrase-pair. Denoting the probability that two words are unaligned as $\bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}) = 1 - p(a_{i,j}|\mathbf{e}, \mathbf{f})$, we can compute, for the inside region, the probability that there is at least one word inside one phrase aligned to a word inside the other phrase as:

$$\alpha(j_1, j_2, i_1, i_2) = 1 - \prod_{(j,i) \in in(j_1, j_2, i_1, i_2)} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}).$$

Similarily for the outside region, we compute the probability that no word inside one phrase is aligned to a word outside the other phrase:

$$\beta(j_1, j_2, i_1, i_2) = \prod_{(j,i) \in out(j_1, j_2, i_1, i_2)} \bar{p}(a_{i,j}|\mathbf{e}, \mathbf{f}).$$

Finally, the same function is used for evaluation and counting ($f_E = f_C$) and defined as the product of these two probabilities:

$$f_C(f_{j_1}^{j_2}, e_{i_1}^{i_2}) = \alpha(j_1, j_2, i_1, i_2) \times \beta(j_1, j_2, i_1, i_2).$$

Alignment Constraints and Selection Criteria Weighted alignment matrices admit flexible alignment constraints and selection criteria. Thresholding enables to better tune the balance between the number of extracted phrase-pairs and the accuracy of their assigned scores. $\mathcal{C}_\mathcal{A}$ requires at least one link inside the phrase-pair to have a probability $p(a_{i,j}|\mathbf{e}, \mathbf{f}) > t_a$. Similar constraints could be applied on links outside the phrase-pair. Likewise, $\mathcal{C}_\mathcal{S}$ admits only phrase-pairs with an evaluation score greater than a threshold $f_E(f_{j_1}^{j_2}, e_{i_1}^{i_2}) > t_p$, and setup a phrase length limit.

Translation Model Scores While the phrase translation probability estimated as ϕ (see step (10) of Algorithm 1) can be applied unchanged to the fractional counts f_C, the lexical scores lex have to be modified to incorporate link probabilities. The main difference is the computation of the lexical probabilities $w(e_i|f_j)$ and $w(f_j|e_i)$, which are calculated using relative occurrence frequencies (Koehn et al., 2003). Instead of simply counting every occurrence once $count(e_i, f_j) = 1$, link probabilities provided by the weighted matrix are used as fractional counts: $count(e_i, f_j) = p(a_{i,j}|\mathbf{e}, \mathbf{f})$ (Liu et al., 2009). Using fractional counts for f_E, f_C and w enables a more accurate evaluation of phrase-pairs depending on the context of the sentence-pair in which they occur, hence a better estimation of their scores.

4 Discriminative Modeling of the Weighted Alignment Matrix

Previous attempts at taking advantage of WAMs have relied on generative alignment models, augmented with some heuristics such as symmetrization have been used to produce the alignment matrices. This approch is less than optimal, since the generative paradigm is not well suited to incorporate arbitrary and possibly interdependent sources of information. Furthermore, all symmetrization heuristics act locally at the sentence-pair level and lack a global view of the training corpus.

To overcome these limitations we propose to view the alignment problem as a structured classification task and model the weighted matrix directly as in (Tomeh et al., 2010). In the presence of manually annotated data with *active* or *inactive* links, a discriminative classifier can be trained to model the probability of each link being *active* using an exponential model:

$$p\left(l_{i,j} = active|\boldsymbol{x}\right) = \frac{\exp \sum_{k=1}^{K} \lambda_k g_k\left(y, \boldsymbol{x}\right)}{Z(\boldsymbol{x})},$$

where $\boldsymbol{x}$ denotes the observation, $Z(\boldsymbol{x})$ is a normalization constant, $(g_k)_{k=1}^{K}$ defines a set of feature functions, and each g_k is associated with a weight λ_k. We use features that describe the linguistic context of a given link, and depend on the sentence pair in which it occurs, augmented by part-of-speech tags and related corpus statistics. We also incorporate the predictions of MGIZA++ alignments as features, which can be viewed as a solution to the symmetrization problem. Since the alignment matrix is typically sparse, with a majority of inactive links, the classification task introduced above is imbalanced. Hence, we only consider the links that occur in the union of all input alignments; all other links are deemed inactive. The model is trained to optimize the log-likelihood of the parameters, regularized using a combination of ℓ^1 and ℓ^2 terms, allowing for efficient feature selection while maintaining numerical stability.

5 Experiments

In our experiments, we aim (1) to compare the standard translation model training method with the method based on weighted alignment matrices; and (2) to contrast different approches to populate the matrices with link posterior probabilities.

For this purpose we build several phrase-based, Arabic to English, translation systems using Moses[1] in its default configuration. In order to tune the parameters of the translation systems, Minimum Error-Rate Training (Och, 2003) is applied on the development corpus, for which we used the NIST MT'06 evaluation's test set, containing 1,797 Arabic sentences (46K words) with four English references (53K words). The performance of each system is assessed by calculating the multi-reference BLEU on NIST MT'08 evaluation's test set, which contains 1,360 Arabic sentences (43K words), each with four references (53K words). For training the various models used by the translation systems, we select a subset of the LDC resources made available by the NIST MT'09 constrained track[2]. In order to validate the obtained results on training corpora of varying sizes, we consider two training conditions, one with 30K parallel sentence pairs, and another with 130K. For each condition, we report below the AER, the BLEU scores on the test set, along with the size of the obtained phrase tables. A 4-gram back-off language model, estimated with SRILM[3] is trained on the NIST MT'09 constrained English data. All Arabic sentences are pre-processed using MADA+TOKAN[4] (Habash and Rambow, 2005), and segmented according to the D2 tokenization scheme. The IBM Arabic-English Word Alignment Corpus (Ittycheriah et al., 2006) is used to train both CRF and MaxEnt aligners and evaluate them using Alignment Error Rate (AER).

5.1 Translation Models Construction

In section 3, we have described a generic algorithm that constructs the translation model in three steps: word alignment, phrase-pairs extraction, and scoring. In this section, we compare different instantiations of these steps, and report the translation performance of the resulting models.

In the word alignment step, we experiment two configurations of the alignment matrix: (i) a standard alignment matrix , which contains the links of the 1-best alignment; and (ii) the weighted alignment matrix, which is populated with link probabilities. Note that we can obtain a matrix in configuration (i) by thresholding the probabilities in the weighted matrix according to a threshold t_a[5]. Hence, for each word aligner (briefly described be-

[1] http://www.statmt.org/moses/
[2] http://www.itl.nist.gov/iad/mig/tests/mt/2009/
[3] http://www-speech.sri.com/projects/srilm/.
[4] http://www1.ccls.columbia.edu/ cadim/MADA.html
[5] In our experiments t_a is set to 0.5.

low) that produce a weighted matrix, we derive two systems: *standard* and *WAM-based* [6].

The two remaining steps depend on the form of the alignment matrix computed in the first step. For standard matrices (i) we use the standard heuristic for extraction, and relative frequencies for scoring (Koehn et al., 2003). For weighted matrices (ii), a phrase posterior $f_C(f_{j1}^{j2}, e_{i1}^{i2})$ can be calculated and used as a fractional count. Only phrase-pairs with a fractional count above certain threshold t_p[7] are extracted. The same fractional counts are used for scoring with relative fractional frequencies. In both configurations, only phrase-pairs that do not exceed a length limit of 7, on the source or the target side, are retained and scored.

5.2 Results and Discussion

In this section, we describe five alignment systems and compare their performance (see Table 5.2).

MGIZA++[8] These alignments are produced by the multi-threaded and optimized alignment toolkit MGIZA++ (Gao and Vogel, 2008), which implements the IBM models. This tool only outputs deterministic alignment matrices in configuration (i). These models also deliver features for the discriminative word aligners described below.

MGIZA++ IBM4 represents the performance of the standard baseline: one IBM4 alignment in each direction, which are symmetrized with *grow-diag-final-and* heuristic. This system deliver competitive BLEU scores of 35.9 and 40.2 on the 30K and 130K respectively, with a much smaller phrase table than all the other systems.

N-best WAM[9] These alignments build weighted matrices, by averaging link occurences over MGIZA++ N-best alignments produced by the IBM model 4, as described in (Liu et al., 2009).

This method slightly improves performance over the baseline. Gains of 0.3 BLEU point on the small task and 0.2 on the larger one are obtained. Improvements are only obtained in the weighted matrix configuration, while standard alignments obtained by thresholding the (10-best based) weighted matrix seem to hurt performance for the selected threshold (0.5). Phrase tables obtained from these systems are only slightly larger than the baseline, which might explain the small improvement. The N-best system achieves comparable AER to the MGIZA++ baseline.

PostCAT[10] The Posterior Constrained Alignment Toolkit (Graça et al., 2007) implements an efficient and principled way to inject rich constraints on the posteriors of latent variables into the EM algorithm, allowing it to satisfy additional, otherwise intractable constraints. When applying constraints such as a *symmetry* or *bijectivity* on a regular HMM alignment, it delivers models that are comparable in accuracy to IBM4 model, and under which statistics to estimate posteriors can still be collected efficiently. This allow us to construct weighted matrices with posteriors estimated over constrained HMM models, by calculating for each link, the average of the posterior given by two HMM models in both translation direction, a method referred to as the *soft union* symmetrization. Our experiments use Geppetto[11] (Ling et al., 2010), an implementation of the weighted alignment matrix integrated with PostCAT.

For the small task, both bijective and symmetric PostCAT alignments, in the standard configuration, outperform MGIZA++ and N-best WAM by ≈ 0.8 BLEU point. The weighted matrix configuration performs even better than the standard one and increases BLEU scores by another ≈ 0.3 BLEU point. Improvements are persistent but less apparent on the larger task. We notice that the phrase table extracted from the weighted matrix is considerably larger than the standard one (by a factor of at least 3). PostCAT also slightly decreases the AER as compared to the MGIZA++ baseline.

CRF[12] The alignment matrix is modeled with a conditional random field (CRF), of which the graphical structure is quite complex and contains many loops (Niehues and Vogel, 2008). Therefore, neither training nor inference can be performed exactly, and the loopy belief propagation algorithm is used to approximate the posteriors. The CRF approach differs from our MaxEnt model (Section 4) in two aspects: first, MaxEnt training only optimizes the log-likelihood, whereas CRF training also aims at minimizing the AER. Second, while both models use the same set of features, MaxEnt

[6] Our experiments show that post-processing the weighted matrix to nullify all link probabilities, that are inferior to a threshold t_a, improves the performance. We use $t_a = 0.5$.

[7] In our experiments t_p is set to 0.1.

[8] http://geek.kyloo.net/

[9] http://www.nlp.org.cn/ liuyang/wam/wam.html

[10] http://www.seas.upenn.edu/ strctlrn/CAT/CAT.html

[11] http://code.google.com/p/geppetto/

[12] We thank J. Niehues (KIT) for sharing his implementation.

	Translation task:		30K					130K				
	Translation model construction:			Standard(i)		WAM(ii)			Standard(i)		WAM(ii)	
	Alignment		**AER**	**BLEU**	**PT**	**BLEU**	**PT**	**AER**	**BLEU**	**PT**	**BLEU**	**PT**
Generative	**MGIZA++**	HMM	28.35	35.01	3,6	-	-	26.77	39.15	9,7	-	-
		IBM4	24.97	35.90	2,4	-	-	23.30	40.18	6,5	-	-
	10-**best**	IBM4	24.92	35.78	2,4	36.21	3,0	23.26	40.00	6,6	40.43	8,5
	PostCAT	Bijective	22.53	36.62	3,3	36.94	10,2	20.49	40.08	9,1	40.61	29,5
		Symmetric	22.48	36.69	2,9	36.96	10,7	20.83	40.24	8,5	40.43	30,2
Discriminative	**CRF**	HMM	25.39	35.93	4,6	36.50	11,9	23.65	39.56	12,6	40.00	31,2
		IBM4	23.51	36.07	3,4	36.93	8,4	22.04	40.34	8,7	40.32	21,3
		HMM+IBM 1,3,4	21.03	36.34	3,7	37.10	8,4	19.65	40.14	9,8	40.35	21,3
	MaxEnt	HMM	17.61	36.90	6,7	37.48	11,7	16.42	40.47	17,7	40.84	30,0
		IBM4	15.61	37.17	5,5	37.52	9,6	14.32	41.04	14,5	41.13	25,0
		HMM+IBM 1,3,4	14.69	37.12	5,2	37.92	8,6	13.92	40.82	13,4	41.08	22,2

Table 1: Comparison of five word aligners: MGIZA++, 10-best, PostCAT, CRF and MaxEnt, in terms of AER, BLEU scores and Phrase Table size in millions (PT). We compare the standard to the WAM-based instantiation of Algorithm 1. Two training corpus of different sizes (30K / 100K) are considered.

turns real-valued features into discrete ones using unsupervised equal frequency interval binning.

On the small task, the CRF approach achieves improvement up to ≈ 0.4 over the MGIZA++ baseline and up to ≈ 1.2 over the WAM-based baseline. Using several input alignments as local features seems beneficial: approximatly 0.5 BLEU point, for both configurations, is gained when using IBM3 and IBM4 features. Similar tendencies are observed for the larger task, albeit with smaller gains. The performance of CRF is comparable to that of PostCAT, but its translation models are however somewhat smaller. Even though the CRF model is trained to maximize the log-likelihood of the manual alignment and to minimize its AER, it achieves only modest improvements in AER over MGIZA++ and PostCAT.

MaxEnt This is the system of Section 4. Discriminative weighted matrices significantly outperform all the previous baselines in both configurations. For the 30K task and for the standard configuration (i): when using only MGIZA++ HMM alignments as input to MaxEnt, we get 1 BLEU point improvement over the standard MGIZA++ IBM4 baseline, and 0.2 point over PostCAT. The extracted phrase table is twice as large as the ones used by the MGIZA++, 10-best or PostCAT. Further improvements are obtained when using IBM4 as input or combining several input alignments, 1.3 BLEU point over MGIZA++ and 0.5 point over PostCAT. MaxEnt based matrices, in configuration (ii), achieve up to 2 BLEU point improvement over MGIZA++ IBM4 and up to 1 point over the best weighted matrix baseline (PostCAT). It is notable that this later improvement is obtained with a smaller phrase table ($\approx$ 25% smaller).

These gains persist for the larger task: MaxEnt in standard (i) configuration is 0.8 BLEU point better than MGIZA++ IBM4, and 0.6 better than PostCAT. In the weighted matrix configuration (ii), these improvements allow us to outperform MGIZA++/IBM4 by nearly 1 BLEU point, 10-best by and approximately 0.7 point, and PostCat by 0.5 point. As for the size of the phrase table, MaxEnt uses smaller phrase tables (22,2M) than PostCAT (30,2M), but much larger ones than MGIZA++ IBM4 (6,5M). Unlike all the other systems, MaxEnt drastically decreases the AER, and achieves approximately 40% relative reduction over MGIZA++ on both 30K and 130K tasks.

6 Conclusion

In this paper we presented a generic algorithm to construct the translation model from a parallel corpus, for which we described two instantiations: standard and WAM-based. We compared several generative and discriminative word aligners in both instantiations, and showed that the WAM-based outperforms the standard procedure due to its improved use of the word alignment probability distribution as compared to the Viterbi alignments. We proposed a discriminative estimation scheme for the probabilities in the weighted matrix using an exponential model and showed that significant

improvements in BLEU scores can be achieved. Our MaxEnt modeling of the matrix led to approximately 2 BLEU points improvement over the standard MGIZA++ baseline, using a small training corpus and 1 BLEU point using a larger one. It is finally interesting to see that, contrarily to the standard training regime, WAM-based training seems to benefit from alignments with better AERs.

7 Acknowledgments

This work was partly realized as part of the Quaero Program, funded by OSEO, the French agency for innovation.

References

Ayan, Necip Fazil and Bonnie J. Dorr. 2006. A maximum entropy approach to combining word alignments. In *Proc. of NAACL-HLT*, pages 96–103.

Birch, Alexandra, Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proc. of WMT*, pages 154–157, New York, NY.

Brown, P. F., V. J. D. Pietra, S. A. D. Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19(2):263–311.

de Gispert, Adrià, Juan Pino, and William Byrne. 2010. Hierarchical phrase-based translation grammars extracted from alignment posterior probabilities. In *Proc. of EMNLP*, pages 545–554.

DeNero, John and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proc. of ACL'08: HLT*, pages 25–28, Columbus, Ohio, June.

DeNero, John, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proc. of WMT*, pages 31–38, New York City.

Deng, Yonggang and William Byrne. 2005. HMM word and phrase alignment for statistical machine translation. In *Proc. of the EMNLP*, pages 169–176.

Galley, M., J. Graehl, K. Knight, D. Marcu, S. DeNeefe, W. Wang, and I. Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proc. of the 21st ICCL and 44th ACL*, pages 961–968, Sydney, Australia.

Gao, Qin and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *SETQA-NLP '08*, pages 49–57.

Graça, João, Kuzman Ganchev, and Ben Taskar. 2007. Expectation maximization and posterior constraints. In *NIPS*.

Habash, Nizar and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proc. of the 43rd ACL*, pages 573–580.

Ittycheriah, Abe, Yasser Al-Onaizan, and Salim Roukos. 2006. The IBM Arabic-English Word Alignment Corpus. Technical report.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. NAACL-HLT 2003*, pages 48–54.

Ling, Wang, Tiago Luís, Joao Graça, Luísa Coheur, and Isabel Trancoso. 2010. Towards a General and Extensible Phrase-Extraction Algorithm. In *Proc. of 7th IWSLT*, pages 313–320.

Liu, Yang, Tian Xia, Xinyan Xiao, and Qun Liu. 2009. Weighted alignment matrices for statistical machine translation. In *Proc. of EMNLP*, pages 1017–1026.

Marcu, Daniel and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. of EMNLP*, pages 133–139.

Mi, Haitao and Liang Huang. 2008. Forest-based translation rule extraction. In *Proc. of the 2008 EMNLP*, pages 206–214, Honolulu, Hawaii.

Niehues, Jan and Stephan Vogel. 2008. Discriminative word alignment via alignment matrix modeling. In *Proc. of WMT*, pages 18–25.

Och, Franz Josef. 2003. Minimum error rate training in statistical machine translation. In *Proc. of the 41st Annual Meeting on ACL*, pages 160–167.

Tomeh, Nadi, Alexandre Allauzen, Guillaume Wisniewski, and François Yvon. 2010. Refining word alignment with discriminative training. In *Proc. of AMTA*, Denver, CO.

Tomeh, Nadi, Thomas Lavergne, Allexandre Allauzen, and François Yvon. 2011. Designing an improved discriminative word aligner. In *Proc. of CICLing*, Tokyo, Japan.

Venugopal, Ashish, Stephan Vogel, and Alex Waibel. 2003. Effective phrase translation extraction from alignment models. In *Proc. of the 41st Annual Meeting of the ACL*, pages 319–326.

Venugopal, Ashish, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2008. Wider pipelines: N-best alignments and parses in MT training. In *Proc. of AMTA*, pages 192–201.

Wang, Wei, Kevin Knight, and Daniel Marcu. 2007. Binarizing syntax trees to improve syntax-based machine translation accuracy. In *Proc. of EMNLP-CoNLL*, pages 746–754, Prague, Czech Republic.

Xue, Yong-Zeng, Sheng Li, Tie-Jun Zhao, Mu-Yun Yang, and Jun Li. 2006. Bilingual phrase extraction from n-best alignments. *ICICIC*, pages 410–414.