Getting the structure right for word alignment: LEAF

Alexander Fraser
ISI / University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
fraser@isi.edu

Daniel Marcu
ISI / University of Southern California
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
marcu@isi.edu

Abstract

Word alignment is the problem of annotating parallel text with translational correspondence. Previous generative word alignment models have made structural assumptions such as the 1-to-1, 1-to-N, or phrase-based consecutive word assumptions, while previous discriminative models have either made such an assumption directly or used features derived from a generative model making one of these assumptions. We present a new generative alignment model which avoids these structural limitations, and show that it is effective when trained using both unsupervised and semi-supervised training methods.

1 Introduction

Several generative models and a large number of discriminatively trained models have been proposed in the literature to solve the problem of automatic word alignment of bitexts. The generative proposals have required unrealistic assumptions about the structure of the word alignments. Two assumptions are particularly common. The first is the 1-to-N assumption, meaning that each source word generates zero or more target words, which requires heuristic techniques in order to obtain alignments suitable for training a SMT system. The second is the consecutive word-based "phrasal SMT" assumption. This does not allow gaps, which can be used to particular advantage by SMT models which model hierarchical structure. Previous discriminative models have either made such assumptions directly or used features from a generative model making such an assumption. Our objective is to automatically produce alignments which can be used to build high quality machine translation systems. These are presumably close to the alignments that trained bilingual speakers produce. Human annotated alignments often contain M-to-N alignments, where several source words are aligned to several target words and the resulting unit can not be further decomposed. Source or target words in a single unit are sometimes nonconsecutive.

In this paper, we describe a new generative model which directly models M-to-N non-consecutive word alignments. The rest of the paper is organized as follows. The generative story is presented, followed by the mathematical formulation. Details of the unsupervised training procedure are described. The generative model is then decomposed into feature functions used in a log-linear model which is trained using a semi-supervised algorithm. Experiments show improvements in word alignment accuracy and usage of the generated alignments in hierarchical and phrasal SMT systems results in an increased BLEU score. Previous work is discussed and this is followed by the conclusion.

2 LEAF: a generative word alignment model

2.1 Generative story

We introduce a new generative story which enables the capture of non-consecutive M-to-N alignment structure. We have attempted to use the same labels as the generative story for Model 4 (Brown et

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 51–60, Prague, June 2007.

al., 1993), which we are extending.

Our generative story describes the stochastic generation of a target string f (sometimes referred to as the French string, or foreign string) from a source string e (sometimes referred to as the English string), consisting of l words. The variable m is the length of f. We generally use the index i to refer to source words (e_i is the English word at position i), and j to refer to target words.

Our generative story makes the distinction between different types of source words. There are head words, non-head words, and deleted words. Similarly, for target words, there are head words, non-head words, and spurious words. A head word is linked to zero or more non-head words; each non-head word is linked to from exactly one head word. The purpose of head words is to try to provide a robust representation of the semantic features necessary to determine translational correspondence. This is similar to the use of syntactic head words in statistical parsers to provide a robust representation of the syntactic features of a parse sub-tree.

A minimal translational correspondence consists of a linkage between a source head word and a target head word (and by implication, the non-head words linked to them). Deleted source words are not involved in a minimal translational correspondence, as they were "deleted" by the translation process. Spurious target words are also not involved in a minimal translational correspondence, as they spontaneously appeared during the generation of other target words.

Figure 1 shows a simple example of the stochastic generation of a French sentence from an English sentence, annotated with the step number in the generative story.

1. Choose the source word type.

   for each $i = 1, 2, ..., l$ choose a word type $\chi_i = -1$ (non-head word), $\chi_i = 0$ (deleted word) or $\chi_i = 1$ (head word) according to the distribution $g(\chi_i|e_i)$

   let $\chi_0 = 1$

2. Choose the identity of the head word for each non-head word.

   for each $i = 1, 2, ..., l$ if $\chi_i = -1$ choose a "linked from head word" value μ_i (the position of the head word which e_i is linked to) according to the distribution $w_{-1}(\mu_i - i|\text{class}_e(e_i))$

   for each $i = 1, 2, ..., l$ if $\chi_i = 1$ let $\mu_i = i$

   for each $i = 1, 2, ..., l$ if $\chi_i = 0$ let $\mu_i = 0$

   for each $i = 1, 2, ..., l$ if $\chi_{\mu_i} \neq 1$ return "failure"

3. Choose the identity of the generated target head word for each source head word.

   for each $i = 1, 2, ..., l$ if $\chi_i = 1$ choose τ_{i1} according to the distribution $t_1(\tau_{i1}|e_i)$

4. Choose the number of words in a target cept conditioned on the identity of the source head word and the source cept size (γ_i is 1 if the cept size is 1, and 2 if the cept size is greater).

   for each $i = 1, 2, ..., l$ if $\chi_i = 1$ choose a Foreign cept size ψ_i according to the distribution $s(\psi_i|e_i, \gamma_i)$

   for each $i = 1, 2, ..., l$ if $\chi_i < 1$ let $\psi_i = 0$

5. Choose the number of spurious words.

   choose ψ_0 according to the distribution $s_0(\psi_0|\sum_i \psi_i)$

   let $m = \psi_0 + \sum_{i=1}^{l} \psi_i$

6. Choose the identity of the spurious words.

   for each $k = 1, 2, ..., \psi_0$ choose τ_{0k} according to the distribution $t_0(\tau_{0k})$

7. Choose the identity of the target non-head words linked to each target head word.

   for each $i = 1, 2, ..., l$ and for each $k = 2, 3, ..., \psi_i$ choose τ_{ik} according to the distribution $t_{>1}(\tau_{ik}|e_i, \text{class}_h(\tau_{i1}))$

8. Choose the position of the target head and non-head words.

   for each $i = 1, 2, ..., l$ and for each $k = 1, 2, ..., \psi_i$ choose a position π_{ik} as follows:

   - if $k = 1$ choose π_{i1} according to the distribution $d_1(\pi_{i1} - c_{\rho_i}|\text{class}_e(e_{\rho_i}), \text{class}_f(\tau_{i1}))$
   - if $k = 2$ choose π_{i2} according to the distribution $d_2(\pi_{i2} - \pi_{i1}|\text{class}_f(\tau_{i1}))$

source	absolutely	[comma]	they	do	not	want	to	spend	that	money
word type (1)	DEL.	DEL.	HEAD	non-head	HEAD	HEAD	non-head	HEAD	HEAD	HEAD
linked from (2)			THEY	do	NOT	WANT	to	SPEND	THAT	MONEY
head(3)			ILS		PAS	DESIRENT		DEPENSER	CET	ARGENT
cept size(4)			1		2	1		1	1	1
num spurious(5)	1									
spurious(6)	aujourd'hui									
non-head(7)			ILS	PAS	ne	DESIRENT	DEPENSER	CET	ARGENT	
placement(8)	aujourd'hui		ILS	ne	DESIRENT	PAS	DEPENSER	CET	ARGENT	
spur. placement(9)			ILS	ne	DESIRENT	PAS	DEPENSER	CET	ARGENT	aujourd'hui

Figure 1: Generative story example, (number) indicates step number

- if $k > 2$ choose π_{ik} according to the distribution $d_{>2}(\pi_{ik} - \pi_{ik-1}|\text{class}_f(\tau_{i1}))$

  if any position was chosen twice, return "failure"

9. Choose the position of the spuriously generated words.

   for each $k = 1, 2, ..., \psi_0$ choose a position π_{0k} from $\psi_0 - k + 1$ remaining vacant positions in $1, 2, ..., m$ according to the uniform distribution

   let f be the string $f\pi_{ik} = \tau_{ik}$

We note that the steps which return "failure" are required because the model is deficient. Deficiency means that a portion of the probability mass in the model is allocated towards generative stories which would result in infeasible alignment structures. Our model has deficiency in the non-spurious target word placement, just as Model 4 does. It has additional deficiency in the source word linking decisions. (Och and Ney, 2003) presented results suggesting that the additional parameters required to ensure that a model is not deficient result in inferior performance, but we plan to study whether this is the case for our generative model in future work.

Given e, f and a candidate alignment a, which represents both the links between source and target head-words and the head-word connections of the non-head words, we would like to calculate $p(f, a|e)$. The formula for this is:

$$\begin{aligned}
p(f,a|e) =& [\prod_{i=1}^{l} g(\chi_i|e_i)] \\
& [\prod_{i=1}^{l} \delta(\chi_i,-1) w_{-1}(\mu_i - i|\text{class}_e(e_i))] \\
& [\prod_{i=1}^{l} \delta(\chi_i,1) t_1(\tau_{i1}|e_i)] \\
& [\prod_{i=1}^{l} \delta(\chi_i,1) s(\psi_i|e_i,\gamma_i)] \\
& [s_0(\psi_0|\sum_{i=1}^{l}\psi_i)] \\
& [\prod_{k=1}^{\psi_0} t_0(\tau_{0k})] \\
& [\prod_{i=1}^{l}\prod_{k=2}^{\psi_i} t_{>1}(\tau_{ik}|e_i,\text{class}_h(\tau_{i1}))] \\
& [\prod_{i=1}^{l}\prod_{k=1}^{\psi_i} D_{ik}(\pi_{ik})]
\end{aligned}$$

where:

$\delta(i, i')$ is the Kronecker delta function which is equal to 1 if $i = i'$ and 0 otherwise.

ρ_i is the position of the closest English head word to the left of the word at i or 0 if there is no such word.

$\text{class}_e(e_i)$ is the word class of the English word at position i, $\text{class}_f(f_j)$ is the word class of the French word at position j, $\text{class}_h(f_j)$ is the word class of the French head word at position j.

p_0 and p_1 are parameters describing the probability of not generating and of generating a target spurious word from each non-spurious target word, $p_0 + p_1 = 1$.

$$m' = \sum_{i=1}^{l} \psi_i \tag{1}$$

$$s_0(\psi_0|m') = \binom{m'}{\psi_0} p_0^{m'-\psi_0} p_1^{\psi_0} \tag{2}$$

$$D_{ik}(j) = \begin{cases} d_1(j - c_{\rho_i}|\text{class}_e(e_{\rho_i}), \text{class}_f(\tau_{ik})) \\ \qquad \text{if } k = 1 \\ d_2(j - \pi_{i1}|\text{class}_f(\tau_{ik})) \\ \qquad \text{if } k = 2 \\ d_{>2}(j - \pi_{ik-1}|\text{class}_f(\tau_{ik})) \\ \qquad \text{if } k > 2 \end{cases} \tag{3}$$

$$\gamma_i = \min(2, \sum_{i'=1}^{l} \delta(\mu_{i'}, i)) \tag{4}$$

$$c_i = \begin{cases} \text{ceiling}(\sum_{k=1}^{\psi_i} \pi_{ik}/\psi_i) & \text{if } \psi_i \neq 0 \\ 0 & \text{if } \psi_i = 0 \end{cases} \tag{5}$$

The alignment structure used in many other models can be modeled using special cases of this framework. We can express the 1-to-N structure of models like Model 4 by disallowing $\chi_i = -1$, while for 1-to-1 structure we both disallow $\chi_i = -1$ and deterministically set $\psi_i = \chi_i$. We can also specialize our generative story to the consecutive word M-to-N alignments used in "phrase-based" models, though in this case the conditioning of the generation decisions would be quite different. This involves adding checks on source and target connection geometry to the generative story which, if violated, would return "failure"; naturally this is at the cost of additional deficiency.

2.2 Unsupervised Parameter Estimation

We can perform maximum likelihood estimation of the parameters of this model in a similar fashion to that of Model 4 (Brown et al., 1993), described thoroughly in (Och and Ney, 2003). We use Viterbi training (Brown et al., 1993) but neighborhood estimation (Al-Onaizan et al., 1999; Och and Ney, 2003) or "pegging" (Brown et al., 1993) could also be used.

To initialize the parameters of the generative model for the first iteration, we use bootstrapping from a 1-to-N and a M-to-1 alignment. We use the intersection of the 1-to-N and M-to-1 alignments to establish the head word relationship, the 1-to-N alignment to delineate the target word cepts, and the M-to-1 alignment to delineate the source word cepts.

In bootstrapping, a problem arises when we encounter infeasible alignment structure where, for instance, a source word generates target words but no link between any of the target words and the source word appears in the intersection, so it is not clear which target word is the target head word. To address this, we consider each of the N generated target words as the target head word in turn and assign this configuration 1/N of the counts.

For each iteration of training we search for the Viterbi solution for millions of sentences. Evidence that inference over the space of all possible alignments is intractable has been presented, for a similar problem, in (Knight, 1999). Unlike phrase-based SMT, left-to-right hypothesis extension using a beam decoder is unlikely to be effective because in word alignment reordering is not limited to a small local window and so the necessary beam would be very large. We are not aware of admissible or inadmissible search heuristics which have been shown to be effective when used in conjunction with a search algorithm similar to A* search for a model predicting over a structure like ours. Therefore we use a simple local search algorithm which operates on complete hypotheses.

(Brown et al., 1993) defined two local search operations for their 1-to-N alignment models 3, 4 and 5. All alignments which are reachable via these operations from the starting alignment are considered. One operation is to change the generation decision for a French word to a different English word (move), and the other is to swap the generation decision for two French words (swap). All possible operations are tried and the best is chosen. This is repeated. The search is terminated when no operation results in an improvement. (Och and Ney, 2003) discussed efficient implementation.

In our model, because the alignment structure is richer, we define the following operations: move French non-head word to new head, move English non-head word to new head, swap heads of two French non-head words, swap heads of two English non-head words, swap English head word links of two French head words, link English word to French word making new head words, unlink English and French head words. We use multiple restarts to try to reduce search errors. (Germann et al., 2004; Marcu and Wong, 2002) have some similar operations without the head word distinction.

3 Semi-supervised parameter estimation

Equation 6 defines a log-linear model. Each feature function h_m has an associated weight λ_m. Given a vector of these weights λ, the alignment search problem, i.e. the search to return the best alignment $\hat{a}$ of the sentences e and f according to the model, is specified by Equation 7.

$$p_\lambda(f, a|e) = \frac{\exp(\sum_m \lambda_m h_m(a, e, f))}{\sum_{a', f'} \exp(\sum_m \lambda_m h_m(a', e, f'))} \tag{6}$$

$$\hat{a} = \operatorname*{argmax}_a \sum_m \lambda_m h_m(f, a, e) \tag{7}$$

We decompose the new generative model presented in Section 2 in both translation directions to provide the initial feature functions for our log-linear model, features 1 to 10 and 16 to 25 in Table 1.

We use backoffs for the translation decisions (features 11 and 26 and the HMM translation tables which are features 12 and 27) and the target cept size distributions (features 13, 14, 28 and 29 in Table 1), as well as heuristics which directly control the number of unaligned words we generate (features 15 and 30 in Table 1).

We use the semi-supervised EMD algorithm (Fraser and Marcu, 2006b) to train the model. The initial M-step bootstraps parameters as described in Section 2.2 from a M-to-1 and a 1-to-N alignment. We then perform the D-step following (Fraser and Marcu, 2006b). Given the feature function parameters estimated in the M-step and the feature function weights λ determined in the D-step, the E-step searches for the Viterbi alignment for the full training corpus.

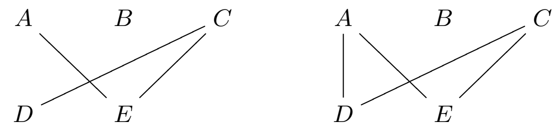

Figure 2: Two alignments with the same translational correspondence

We use $1 -$ F-Measure as our error criterion. (Fraser and Marcu, 2006a) established that it is important to tune α (the trade-off between Precision and Recall) to maximize performance. In working with LEAF, we discovered a methodological problem with our baseline systems, which is that two alignments which have the same translational correspondence can have different F-Measures. An example is shown in Figure 2.

To overcome this problem we fully interlinked the transitive closure of the undirected bigraph formed by each alignment hypothesized by our baseline alignment systems[1]. This operation maps the alignment shown to the left in Figure 2 to the alignment shown to the right. This operation does not change the collection of phrases or rules extracted from a hypothesized alignment, see, for instance, (Koehn et al., 2003). Working with this fully interlinked representation we found that the best settings of α were $\alpha = 0.1$ for the Arabic/English task and $\alpha = 0.4$ for the French/English task.

4 Experiments

4.1 Data Sets

We perform experiments on two large alignments tasks, for Arabic/English and French/English data sets. Statistics for these sets are shown in Table 2. All of the data used is available from the Linguistic Data Consortium except for the French/English

[1] All of the gold standard alignments were fully interlinked as distributed. We did not modify the gold standard alignments.

1	$chi(\chi_i\|e_i)$ source word type	9	$d_2(\triangle j\|\text{class}_f(f_j))$ movement for left-most target non-head word
2	$\mu(\triangle i\|\text{class}_e(e_i))$ choosing a head word	10	$d_{>2}(\triangle j\|\text{class}_f(f_j))$ movement for subsequent target non-head words
3	$t_1(f_j\|e_i)$ head word translation	11	$t(f_j\|e_i)$ translation without dependency on word-type
4	$s(\psi_i\|e_i, \gamma_i)$ ψ_i is number of words in target cept	12	$t(f_j\|e_i)$ translation table from final HMM iteration
5	$s_0(\psi_0\| \sum_i \psi_i)$ number of unaligned target words	13	$s(\psi_i\|\gamma_i)$ target cept size without dependency on source head word e
6	$t_0(f_j)$ identity of unaligned target words	14	$s(\psi_i\|e_i)$ target cept size without dependency on γ_i
7	$t_{>1}(f_j\|e_i, \text{class}_h(\tau_{i1}))$ non-head word translation	15	target spurious word penalty
8	$d_1(\triangle j\|\text{class}_e(e_\rho), \text{class}_f(f_j))$ movement for target head words	16-30	(same features, other direction)

Table 1: Feature functions

gold standard alignments which are available from the authors.

4.2 Experiments

To build all alignment systems, we start with 5 iterations of Model 1 followed by 4 iterations of HMM (Vogel et al., 1996), as implemented in GIZA++ (Och and Ney, 2003).

For all non-LEAF systems, we take the best performing of the "union", "refined" and "intersection" symmetrization heuristics (Och and Ney, 2003) to combine the 1-to-N and M-to-1 directions resulting in a M-to-N alignment. Because these systems do not output fully linked alignments, we fully link the resulting alignments as described at the end of Section 3. The reader should recall that this does not change the set of rules or phrases that can be extracted using the alignment.

We perform one main comparison, which is of semi-supervised systems, which is what we will use to produce alignments for SMT. We compare semi-supervised LEAF with a previous state of the art semi-supervised system (Fraser and Marcu, 2006b). We performed translation experiments on the alignments generated using semi-supervised training to verify that the improvements in F-Measure result in increases in BLEU.

We also compare the unsupervised LEAF system with GIZA++ Model 4 to give some idea of the performance of the unsupervised model. We made an effort to optimize the free parameters of GIZA++, while for unsupervised LEAF there are no free parameters to optimize. A single iteration of unsupervised LEAF[2] is compared with heuristic symmetrization of GIZA++'s extension of Model 4 (which was run for four iterations). LEAF was bootstrapped as described in Section 2.2 from the HMM Viterbi alignments.

Results for the experiments on the French/English data set are shown in Table 3. We ran GIZA++ for four iterations of Model 4 and used the "refined" heuristic (line 1). We ran the baseline semi-supervised system for two iterations (line 2), and in contrast with (Fraser and Marcu, 2006b) we found that the best symmetrization heuristic for this system was "union", which is most likely due to our use of fully linked alignments which was discussed at the end of Section 3. We observe that LEAF unsupervised (line 3) is competitive with GIZA++ (line 1), and is in fact competitive with the baseline semi-supervised result (line 2). We ran the LEAF semi-supervised system for two iterations (line 4). The best result is the LEAF semi-supervised system, with a gain of 1.8 F-Measure over the LEAF unsupervised system.

For French/English translation we use a state of the art phrase-based MT system similar to (Och and Ney, 2004; Koehn et al., 2003). The translation test data is described in Table 2. We use two trigram language models, one built using the English portion of the training data and the other built using additional English news data. The BLEU scores reported in this work are calculated using lowercased and tokenized data. For semi-supervised LEAF the gain of 0.46 BLEU over the semi-supervised baseline is not statistically significant (a gain of 0.78 BLEU would be required), but LEAF semi-supervised compared with GIZA++ is significant, with a gain of 1.23 BLEU. We note that this shows a large gain in trans-

[2]Unsupervised LEAF is equivalent to using the log-linear model and setting $\lambda_m = 1$ for $m = 1$ to 10 and $m = 16$ to 25, while setting $\lambda_m = 0$ for other values of m.

		ARABIC/ENGLISH		FRENCH/ENGLISH	
		A	E	F	E
TRAINING	SENTS	6,609,162		2,842,184	
	WORDS	147,165,003	168,301,299	75,794,254	67,366,819
	VOCAB	642,518	352,357	149,568	114,907
	SINGLETONS	256,778	158,544	60,651	47,765
ALIGN DISCR.	SENTS	1,000		110	
	WORDS	26,882	37,635	1,888	1,726
	LINKS	39,931		2,292	
ALIGN TEST	SENTS	83		110	
	WORDS	1,510	2,030	1,899	1,716
	LINKS	2,131		2,176	
TRANS. DEV	SENTS	728 (4 REFERENCES)		833 (1 REFERENCE)	
	WORDS	18,255	22.0K TO 24.6K	20,562	17,454
TRANS. TEST	SENTS	1,056 (4 REFERENCES)		2,380 (1 REFERENCE)	
	WORDS	28,505	35.8K TO 38.1K	58,990	49,182

Table 2: Data sets

lation quality over that obtained using GIZA++ because BLEU is calculated using only a single reference for the French/English task.

Results for the Arabic/English data set are also shown in Table 3. We used a large gold standard word alignment set available from the LDC. We ran GIZA++ for four iterations of Model 4 and used the "union" heuristic. We compare GIZA++ (line 1) with one iteration of the unsupervised LEAF model (line 2). The unsupervised LEAF system is worse than four iterations of GIZA++ Model 4. We believe that the features in LEAF are too high dimensional to use for the Arabic/English task without the backoffs available in the semi-supervised models. The baseline semi-supervised system (line 3) was run for three iterations and the resulting alignments were combined with the "union" heuristic. We ran the LEAF semi-supervised system for two iterations. The best result is the LEAF semi-supervised system (line 4), with a gain of 5.4 F-Measure over the baseline semi-supervised system.

For Arabic/English translation we train a state of the art hierarchical model similar to (Chiang, 2005) using our Viterbi alignments. The translation test data used is described in Table 2. We use two trigram language models, one built using the English portion of the training data and the other built using additional English news data. The test set is from the NIST 2005 translation task. LEAF had the best performance scoring 1.43 BLEU better than the baseline semi-supervised system, which is statistically significant.

5 Previous Work

The LEAF model is inspired by the literature on generative modeling for statistical word alignment and particularly by Model 4 (Brown et al., 1993). Much of the additional work on generative modeling of 1-to-N word alignments is based on the HMM model (Vogel et al., 1996). (Toutanova et al., 2002) and (Lopez and Resnik, 2005) presented a variety of refinements of the HMM model particularly effective for low data conditions. (Deng and Byrne, 2005) described work on extending the HMM model using a bigram formulation to generate 1-to-N alignment structure. The common thread connecting these works is their reliance on the 1-to-N approximation, while we have defined a generative model which does not require use of this approximation, at the cost of having to rely on local search.

There has also been work on generative models for other alignment structures. (Wang and Waibel, 1998) introduced a generative story based on extension of the generative story of Model 4. The alignment structure modeled was "consecutive M to non-consecutive N". (Marcu and Wong, 2002) defined the Joint model, which modeled consecutive word M-to-N alignments. (Matusov et al., 2004) presented a model capable of modeling 1-to-N and M-to-1 alignments (but not arbitrary M-to-N alignments) which was bootstrapped from Model 4. LEAF directly models non-consecutive M-to-N alignments.

One important aspect of LEAF is its symmetry. (Och and Ney, 2003) invented heuristic symmetriza-

SYSTEM	FRENCH/ENGLISH F-MEASURE ($\alpha = 0.4$)	FRENCH/ENGLISH BLEU	ARABIC/ENGLISH F-MEASURE ($\alpha = 0.1$)	ARABIC/ENGLISH BLEU
GIZA++	73.5	30.63	75.8	51.55
(FRASER AND MARCU, 2006B)	74.1	31.40	79.1	52.89
LEAF UNSUPERVISED	74.5		72.3	
LEAF SEMI-SUPERVISED	76.3	31.86	84.5	54.34

Table 3: Experimental Results

tion of the output of a 1-to-N model and a M-to-1 model resulting in a M-to-N alignment, this was extended in (Koehn et al., 2003). We have used insights from these works to help determine the structure of our generative model. (Zens et al., 2004) introduced a model featuring a symmetrized lexicon. (Liang et al., 2006) showed how to train two HMM models, a 1-to-N model and a M-to-1 model, to agree in predicting all of the links generated, resulting in a 1-to-1 alignment with occasional rare 1-to-N or M-to-1 links. We improve on these works by choosing a new structure for our generative model, the head word link structure, which is both symmetric and a robust structure for modeling of non-consecutive M-to-N alignments.

In designing LEAF, we were also inspired by dependency-based alignment models (Wu, 1997; Alshawi et al., 2000; Yamada and Knight, 2001; Cherry and Lin, 2003; Zhang and Gildea, 2004). In contrast with their approaches, we have a very flat, one-level notion of dependency, which is bilingually motivated and learned automatically from the parallel corpus. This idea of dependency has some similarity with hierarchical SMT models such as (Chiang, 2005).

The discriminative component of our work is based on a plethora of recent literature. This literature generally views the discriminative modeling problem as a supervised problem involving the combination of heuristically derived feature functions. These feature functions generally include the prediction of some type of generative model, such as the HMM model or Model 4. A discriminatively trained 1-to-N model with feature functions specifically designed for Arabic was presented in (Ittycheriah and Roukos, 2005). (Lacoste-Julien et al., 2006) created a discriminative model able to model 1-to-1, 1-to-2 and 2-to-1 alignments for which the best results were obtained using features based on symmetric HMMs trained to agree, (Liang et al., 2006), and intersected Model 4. (Ayan and Dorr, 2006) defined a discriminative model which learns how to combine the predictions of several alignment algorithms. The experiments performed included Model 4 and the HMM extensions of (Lopez and Resnik, 2005). (Moore et al., 2006) introduced a discriminative model of 1-to-N and M-to-1 alignments, and similarly to (Lacoste-Julien et al., 2006) the best results were obtained using HMMs trained to agree and intersected Model 4. LEAF is not bound by the structural restrictions present either directly in these models, or in the features derived from the generative models used. We also iterate the generative/discriminative process, which allows the discriminative predictions to influence the generative model.

Our work is most similar to work using discriminative log-linear models for alignment, which is similar to discriminative log-linear models used for the SMT decoding (translation) problem (Och and Ney, 2002; Och, 2003). (Liu et al., 2005) presented a log-linear model combining IBM Model 3 trained in both directions with heuristic features which resulted in a 1-to-1 alignment. (Fraser and Marcu, 2006b) described symmetrized training of a 1-to-N log-linear model and a M-to-1 log-linear model. These models took advantage of features derived from both training directions, similar to the symmetrized lexicons of (Zens et al., 2004), including features derived from the HMM model and Model 4. However, despite the symmetric lexicons, these models were only able to optimize the performance of the 1-to-N model and the M-to-1 model separately, and the predictions of the two models required combination with symmetrization heuristics. We have overcome the limitations of that work by defining new feature functions, based on the LEAF generative model, which score non-consecutive M-to-N alignments so that the final performance criterion can be optimized directly.

6 Conclusion

We have found a new structure over which we can robustly predict which directly models translational correspondence commensurate with how it is used in hierarchical SMT systems. Our new generative model, LEAF, is able to model alignments which consist of M-to-N non-consecutive translational correspondences. Unsupervised LEAF is comparable with a strong baseline. When coupled with a discriminative training procedure, the model leads to increases between 3 and 9 F-score points in alignment accuracy and 1.2 and 2.8 BLEU points in translation accuracy over strong French/English and Arabic/English baselines.

7 Acknowledgments

This work was partially supported under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022. We would like to thank the USC Center for High Performance Computing and Communications.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John D. Lafferty, I. Dan Melamed, David Purdy, Franz J. Och, Noah A. Smith, and David Yarowsky. 1999. Statistical machine translation, final report, JHU workshop.

Hiyan Alshawi, Srinivas Bangalore, and Shona Douglas. 2000. Learning dependency translation models as collections of finite state head transducers. *Computational Linguistics*, 26(1):45–60.

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

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

Colin Cherry and Dekang Lin. 2003. A probability model to improve word alignment. In *Proceedings of ACL*, pages 88–95, Sapporo, Japan.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL*, pages 263–270, Ann Arbor, MI.

Yonggang Deng and William Byrne. 2005. Hmm word and phrase alignment for statistical machine translation. In *Proceedings of HLT-EMNLP*, Vancouver, Canada.

Alexander Fraser and Daniel Marcu. 2006a. Measuring word alignment quality for statistical machine translation. In *Technical Report ISI-TR-616*, ISI/University of Southern California.

Alexander Fraser and Daniel Marcu. 2006b. Semi-supervised training for statistical word alignment. In *Proceedings of COLING-ACL*, pages 769–776, Sydney, Australia.

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2004. Fast decoding and optimal decoding for machine translation. *Artificial Intelligence*, 154(1-2):127–143.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for Arabic-English machine translation. In *Proceedings of HLT-EMNLP*, pages 89–96, Vancouver, Canada.

Kevin Knight. 1999. Decoding complexity in word-replacement translation models. *Computational Linguistics*, 25(4):607–615.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACL*, pages 127–133, Edmonton, Canada.

Simon Lacoste-Julien, Dan Klein, Ben Taskar, and Michael Jordan. 2006. Word alignment via quadratic assignment. In *Proceedings of HLT-NAACL*, pages 112–119, New York, NY.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of HLT-NAACL*, New York.

Yang Liu, Qun Liu, and Shouxun Lin. 2005. Log-linear models for word alignment. In *Proceedings of ACL*, pages 459–466, Ann Arbor, MI.

Adam Lopez and Philip Resnik. 2005. Improved hmm alignment models for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 83–86, Ann Arbor, MI.

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

Evgeny Matusov, Richard Zens, and Hermann Ney. 2004. Symmetric word alignments for statistical machine translation. In *Proceedings of COLING*, Geneva, Switzerland.

Robert C. Moore, Wen-Tau Yih, and Andreas Bode. 2006. Improved discriminative bilingual word alignment. In *Proceedings of COLING-ACL*, pages 513–520, Sydney, Australia.

Franz J. Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of ACL*, pages 295–302, Philadelphia, PA.

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

Franz J. Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(1):417–449.

Franz J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan.

Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning. 2002. Extensions to hmm-based statistical word alignment models. In *Proceedings of EMNLP*, Philadelphia, PA.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING*, pages 836–841, Copenhagen, Denmark.

Ye-Yi Wang and Alex Waibel. 1998. Modeling with structures in statistical machine translation. In *Proceedings of COLING-ACL*, volume 2, pages 1357–1363, Montreal, Canada.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403.

Kenji Yamada and Kevin Knight. 2001. A syntax-based statistical translation model. In *Proceedings of ACL*, pages 523–530, Toulouse, France.

Richard Zens, Evgeny Matusov, and Hermann Ney. 2004. Improved word alignment using a symmetric lexicon model. In *Proceedings of COLING*, Geneva, Switzerland.

Hao Zhang and Daniel Gildea. 2004. Syntax-based alignment: Supervised or unsupervised? In *Proceedings of COLING*, Geneva, Switzerland.