Submitted to IEEE Transactions on Audio, Speech & Language Processing

HMM Word and Phrase Alignment for Statistical Machine Translation

Yonggang Deng$^{(1)}$, *Member, IEEE*, William Byrne$^{(2)}$, *Member, IEEE*
IBM T. J. Watson Research Center, Yorktown Heights, N.Y. 10598 USA $^{(1)}$
Cambridge University Engineering Dept, Cambridge CB2 1PZ UK $^{(2)}$

Abstract

Efficient estimation and alignment procedures for word and phrase alignment HMMs are developed for the alignment of parallel text. The development of these models is motivated by an analysis of the desirable features of IBM Model 4, one of the original and most effective models for word alignment. These models are formulated to capture the desirable aspects of Model 4 in an HMM alignment formalism. Alignment behavior is analyzed and compared to human-generated reference alignments, and the ability of these models to capture different types of alignment phenomena is evaluated. In analyzing alignment performance, Chinese-English word alignments are shown to be comparable to those of IBM Model-4 even when models are trained over large parallel texts. In translation performance, phrase-based statistical machine translation systems based on these HMM alignments can equal and exceed systems based on Model-4 alignments, and this is shown in Arabic-English and Chinese-English translation. These alignment models can also be used to generate posterior statistics over collections of parallel text, and this is used to refine and extend phrase translation tables with a resulting improvement in translation quality.

I. Introduction

Alignment is one of the central modeling problems in statistical machine translation (SMT). Given a collection of parallel text - text in one language accompanied by its translation in another language - the process of alignment identifies translation equivalence between documents, paragraphs, sentences, and, within sentences, between words and phrases [1], [2], [3], [4], [5], [6], [7]. This work investigates

This work was funded by ONR MURI Grant N00014-01-1-0685.

the use and refinement of Hidden Markov Models for the automatic alignment of words and phrases in parallel text.

The motivation for improved model-based alignment procedures is due to the role of word-aligned parallel text in current phrase-based SMT systems. Typically, parameters of a statistical word alignment model are estimated from parallel text, and the model is used to generate word alignments over the same texts used in parameter estimation. Phrase pairs, which are word sequences whose words align to each other, are extracted from the alignments and form the basis for phrase-based SMT systems. In this approach the quality of the underlying word alignments has a strong influence on phrase-based SMT system performance. The common practice therefore is to extract phrase pairs from the best attainable word alignments. Model-4 alignments [2] as produced by the GIZA++ Toolkit [8] are currently among the best that can be obtained, especially with large parallel texts.

Despite its modeling power and widespread use, Model-4 has shortcomings. Its formulation is such that maximum likelihood parameter estimation and parallel text alignment are implemented by approximate, hill-climbing, algorithms. As a consequence parameter estimation can be slow, memory intensive, and difficult to parallelize. It is also difficult to compute statistics under Model-4, for instance as needed by EM and these models are of limited usefulness for tasks other than the generation of word alignments; for example, it is computationally difficult to build a translation system based directly on Model-4.

Hidden Markov models (HMMs) are potentially an attractive alternative to Model-4 for word alignment [9], [10] and phrase alignment of parallel text. These models, with their alignment and estimation procedures, now define the mainstream of automatic speech recognition (ASR) [11]. Although SMT is a different task from ASR, there are strong connections between them. Both employ the source-channel model, and both translation and transcription can be performed by source-channel decoding algorithms. However, SMT involves more complex alignment problems. Using HMMs in speech recognition is now fairly straightforward since the temporal nature of speech leads naturally to a left-to-right model topology, whereas translation involves the reordering and the long-range movement of words in ways rarely encountered in modeling speech.

In this paper we describe an HMM alignment framework developed as an alternative to Model-4. Our approach was to analyze the strengths of Model-4 and attempt to introduce them into HMM alignment models while still ensuring that the associated parameter estimation and word alignment procedures remain efficient. In the word alignment and phrase-based translation experiments that will be presented, its performance is comparable to or better than Model-4. As practical benefits of this modeling approach, we can train the alignment models by exact implementations of the Forward-Backward algorithm; by parallelizing estimation, we can control memory usage, reduce the time needed for training, and increase the amount of parallel text used for training. We can also compute statistics under these models in ways not feasible with Model-4, and we show the value of this in the extraction of phrase pairs from parallel text under model-based posterior distributions. Aspects of this framework have been presented in preliminary form [12]. We now provide details and derivations of alignment models and algorithms, as well as experimental results and comparisons to other modeling approaches that demonstrate that these techniques are robust and useful.

The paper proceeds as follows. The word-to-phrase alignment HMM which forms basis of the modeling methodology is formally presented in Section II. We describe the basic model components and compare it to IBM word alignment models by contrasting the details of their generative formulation. In particular, we identify their key model components, analyze their weaknesses and strengths, and explain what features from Model 4 we have attempted to introduce into our model. In Section III we discuss the embedded estimation procedures developed for word-to-phrase alignment HMMs and address smoothing issues for robust parameter estimation. We also discuss deriving word alignments under the model and some possible model refinements. In Section IV we discuss word alignment induced statistical phrase translation models with a focus on phrase pair extraction. We present a model-based phrase pair distribution, which enables alternative phrase translation extraction. We show a simple strategy of improving phrase pair extraction by phrase pair posterior, as an example of using the word-to-phrase alignment models, not just the alignments generated over parallel text. In Section V, we investigate word alignment performance and show machine translation evaluation results on Chinese-English and Arabic-English translation systems. We compare the word-to-phrase alignment model with Model-4 in the tasks of word alignment and translation.

II. Word-to-Phrase Hidden Markov Model Alignment

We begin with a detailed description of the component distributions and random variables needed to describe a generative probabilistic model of word-to-phrase alignment.

A. *Component Variables and Distributions*

We start with a source sentence of I words, $\mathbf{s} = s_1^I$, and its translation as a J word sentence in the target language, $\mathbf{t} = t_1^J$. We assume that we have a pair of correct translations whose word alignment is unknown; by ‘correct’ we assume that the sentences to be aligned are in fact translations of each other and that spurious sentence pairs have been filtered from the parallel text. We model the generation of the target language word sequence via an intermediate sequence of target language phrases. Here, ‘phrase’ refers only to variable length word sequences in the target language; any subsequence found in the target language sentence can serve as a phrase.

We introduce only a minimal structure to describe the segmentation of target sentences into phrase sequences. We define the Phrase Count variable K, which specifies that the target language sentence is segmented into a sequence of phrases: $\mathbf{t} = v_1^K$. The central modeling assumption is that each phrase in the target phrase sequence v_1^K is generated as a translation of a single word in the source language sentence. The correspondence between source words and target phrases is determined by the alignment sequence a_1^K. In this way, the k^{th} target phrase is generated by the word appearing in position a_k of the source sentence: $s_{a_k} \rightarrow v_k$. The number of words in each target phrase is specified by the random process ϕ_k. This process is necessarily constrained so that the number of words in the phrase sequence agrees with the target sentence length: $J = \sum_{k=1}^{K} \phi_k$.

It is necessary as a practical matter in modeling translation alignment to allow for the insertion of target phrases. This need arises because the correspondence between source sentence and target sentence is not always exact, despite the assumption of correct sentence-level translations. In some instances it may be better to insert phrases rather than insist that they align to a source word. This is typically done by allowing alignments to a non-existent NULL source word. An alternative formulation is to introduce a binary ‘hallucination’ sequence h_1^K that determines how each phrase is generated: if $h_k = 0$, then NULL $\rightarrow v_k$; if $h_k = 1$ then $s_{a_k} \rightarrow v_k$. If the hallucination process takes a value of 0, the corresponding phrase is hallucinated rather than generated as a translation of one of the words in the source sentence.

Taken together, these quantities describe a phrase segmentation of the target language sentence and its alignment to the source sentence: $\mathbf{a} = (\phi_1^K, a_1^K, h_1^K, K)$. The modeling objective is to define a conditional distribution $P(\mathbf{t}, \mathbf{a}|\mathbf{s})$ over these alignments. With the assumption that $P(\mathbf{t}, \mathbf{a}|\mathbf{s}) = 0$ if $\mathbf{t} \neq v_1^K$, we write $P(\mathbf{t}, \mathbf{a}|\mathbf{s}) = P(v_1^K, K, a_1^K, h_1^K, \phi_1^K|\mathbf{s})$ and

$$\begin{aligned} P(v_1^K, K, a_1^K, h_1^K, \phi_1^K|\mathbf{s}) &= P(K|J, \mathbf{s}) \\ &\times P(a_1^K, \phi_1^K, h_1^K|K, J, \mathbf{s}) \times P(v_1^K|a_1^K, h_1^K, \phi_1^K, K, J, \mathbf{s}). \end{aligned}$$

These are the natural dependencies of the component variables in this generative formulation of the word-to-phrase alignment. We now describe the simplifying assumptions made in their realization. The objective is to not to define the ideal realization of each component; many of the assumptions are admittedly simplistic. However simplicity is preferred wherever possible so as to control the complexity of the overall model.

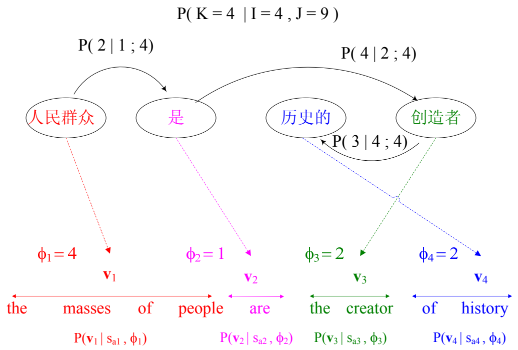

Fig. 1. Simplified Example of Word-to-Phrase HMM Alignment. A Markov network is established by treating source words as Markov states, with the state dependent observation distributions defined over phrases of target words.

a) Phrase Count Distribution: $P(K|J, \mathbf{s})$ specifies the distribution over the number of phrases in the target sentence given the source sentence and the number of words in the target sentence. We use a simple, single parameter distribution $P(K|J, \mathbf{s}) = P(K|J, I) \propto \eta^K$. The scalar $\eta \geq 1$ controls the segmentation of the target sentence into phrases in that larger values of η favor target sentence segmentations with many short phrases. In practice, we use η as a tuning parameter to control the length of the hypothesized target phrases.

b) Word-to-Phrase Alignment Distribution: Before the words of the target language phrases are generated, the alignment of the target phrases to the source words is determined. The alignment is modeled as a Markov process that specifies the lengths of phrases and the alignment of each to one of the source word positions

$$\begin{aligned}
& P(a_1^K, h_1^K, \phi_1^K | K, J, \mathbf{s}) \\
& \quad = \prod_{k=1}^{K} P(a_k, h_k, \phi_k | a_{k-1}, \phi_{k-1}, h_{k-1}, K, J, \mathbf{s}) \\
& \quad = \prod_{k=1}^{K} p(a_k | a_{k-1}, h_k; I) \cdot d(h_k) \cdot n(\phi_k; s_{a_k})
\end{aligned}$$

The actual word-to-phrase alignment (a_k) is a Markov process over the source sentence word indices, as in word-to-word HMM alignment [9]. It is formulated with a dependency on the hallucination variable so that target phrases can be inserted without disrupting the Markov dependencies of phrases aligned to non-NULL source words

$$p(a_j|a_{j-1},h_j;I) = \begin{cases} 1 & a_j = a_{j-1},\ h_j = 0 \\ 0 & a_j \neq a_{j-1},\ h_j = 0 \\ p_a(a_j|a_{j-1};I) & h_j = 1 \end{cases}$$

The target phrase length model $n(\phi;s)$ is a form of source word fertility [2]. It specifies the probability that a source word s generates a target phrase of ϕ words. A distribution $n(\phi;s)$ over the values $\phi = 1,\cdots,N$ is maintained as a table for each source word. The model also requires a table of Markov transition probabilities $p_a(i'|i;I)$ for all source sentence lengths I.

The hallucination sequence is a simple i.i.d. process, where $d(0) = p_0$ and $d(1) = 1 - p_0$. Specified in this way, p_0 acts as a tuning parameter that controls the tendency towards the insertion of target phrases.

c) Word-to-Phrase Translation: The translation of words to phrases is given as

$$P(v_1^K|a_1^K,h_1^K,\phi_1^K,K,J,\mathbf{s}) = \prod_{k=1}^{K} p(v_k|s_{a_k},h_k,\phi_k)$$

so that target phrases are conditionally independent given the individual source words. We define two models of word-to-phrase translation.

The simplest model of word-to-phrase translation is based on context-independent, word-to-word translation: target phrase words are translated independently from the source word via fixed translation tables: $p(v_k|s_{a_k},h_k,\phi_k) = \prod_{j=1}^{\phi_k} t_1(v_k[j]\,|\,h_k \cdot s_{a_k})$ where the notation $h_k \cdot s_{a_k}$ is shorthand for

$$h_k \cdot s_{a_k} = \begin{cases} s_{a_k} & h_k = 1 \\ \text{NULL} & h_k = 0 \end{cases}.$$

In this way specialized translation tables can be maintained for hallucinated phrases to allow their statistics to differ from phrases that arise from direct translation of specific source words.

A more complex realization of word-to-phrase translation captures word context within the target language phrase via *bigram translation probabilities*

$$p(v_k|s_{a_k},h_k,\phi_k) = t_1(v_k[1]\,|\,h_k \cdot s_{a_k}) \times \prod_{j=2}^{\phi_k} t_2(v_k[j]\,|\,v_k[j-1],h_k \cdot s_{a_k})$$

Here, $t_1(t|s)$ is the usual context independent word-to-word translation probability described initially. The bigram translation probability $t_2(t|t',s)$ specifies the likelihood that target word t is to follow t' in a phrase generated by source word s. Note that the conditioning is on words within the target phrase; this respects the conditional independence assumptions and is computationally tractable.

To summarize, the parameter set θ of this formulation of the word-to-phrase alignment HMM consists of the Markov transition matrices p_a, the phrase length tables n, the hallucination parameter p_0, the unigram word-to-word translation table t_1, and the bigram translation probabilities t_2: $\theta = \{p_a(i|i'; I), n(\phi; s), p_0, t_1(t|s), t_2(t|t', s)\}$

The stochastic process by which a source string $\mathbf{s}$ generates a target string $\mathbf{t}$ of J words is summarized as follows (see also Fig. 1):

1) The number of phrases in the target sentence is chosen under $P(K|I, J)$.
2) For each of the K target phrases to be produced :
   a) The alignment a_1^K is generated along with the hallucination process h_1^K.
   b) With the alignment of the k^{th} phrase to the a_k^{th} source word set, the number of words in the k^{th} phrase is then chosen under distribution $n(\phi_k; s_{a_k})$. The ϕ_k satisfy $\sum_{k=1}^{K} \phi_k = J$.
   c) The words in the target phrase v_k are chosen under $P(v_k|s_{a_k}, h_k, \phi_k)$, where the hallucination process controls the insertion of target phrases.
3) The target sentence is formed from the target phrase sequence: $\mathbf{t} = v_1^K$.

Although it incorporates target phrases, the word-to-phrase alignment HMM is very much a model of word translation in that target phrases are produced as sequences of words generated from a single source word. Phrase-level information is used primarily to influence the translation of individual words. The alignment of source and target words can easily be derived from the word-to-phrase alignments: words in a target phrase are aligned to the source word that generated the phrase.

B. Models of Word and Phrase Alignment in Translation

There is extensive prior work in alignment and translation that incorporates phrases, although, to our knowledge, the word-to-phrase HMM is the first non-syntactic model of alignment that directly aligns words to phrases. We now review prior work on word alignment and discuss its influence on this formulation of the word-to-phrase alignment HMM.

1) Hidden Markov Models of Word Alignment: The original formulation of the HMM word alignment model [9] associates each source word s_i with a state in a Markov process. The target sentence t_1^J is the HMM observation sequence in which target words are emitted one-by-one, in target language order, with each state transition :

$$P(\mathbf{t}, \mathbf{a}|\mathbf{s}) = P(\mathbf{a}|\mathbf{s})\ P(\mathbf{t}|\mathbf{s}, \mathbf{a}) = \prod_{j=1}^{J} P(a_j|a_{j-1}; I)\ t(t_j|s_{a_j})\ .$$

The relationship between the word-to-phrase alignment HMM and the original word-to-word HMM alignment model is straightforward: constraining the phrase length component $n(\phi; s)$ to permit only sin-

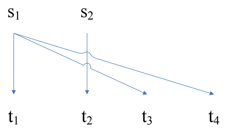

Fig. 2. An Example Alignment with Word-to-Word and Word-to-Phrase Links

gle word phrases reduces the word-to-phrase alignment HMM to a model of word-to-word alignment. We note that the word-to-phrase alignment HMM is very similar to segmental Hidden Markov Models [13], [14], in which HMM states emit observation sequences or trajectories rather than individual observations. We note also that the hallucination process is motivated by the use of NULL alignments in the Markov alignment models [7], although the formulation here is different.

2) Phrase Length Distributions: Developing a model of the production of phrase sequences rather than word sequences required the specification of the phrase count and phrase length models. The form of the phrase length model presented here was motivated by the use of 'stay' probabilities in HMM word-to-word alignment [10], which are applied to encourage the alignment process to remain in a source word state while generating successive target words. By comparison, the Word-to-Phrase HMM alignment models contain detailed models of state occupancy that are more powerful than a single 'stay' parameter.

Fig. 2 gives an example in which a single source word s_1 generates the target words t_3 and t_4. There are multiple alignment sequences which could be associated with this set of links, but more importantly, the target words t_3 and t_4 could be generated either as two one-word phrases or as a single two-word phrase. The balance between word-to-word and word-to-phrase alignments is set by the phrase count distribution parameter η. As η increases, alignments with shorter phrases are favored, whereas for very large η the model effectively permits only word-to-word alignments. As will be shown in Section V-B, it is desirable to have a balanced distribution of word-to-word and word-to-phrase links to obtain the best overall word alignment quality, and η serves as a tuning parameter.

3) Bigram Translation Probabilities: Modeling the production of phrase sequences rather than word sequences also allows the introduction of the bigram translation distribution, which are motivated by statistical techniques developed for word sense disambiguation in translation (e.g. [15], [16]). Here, the likelihood assigned to the words in a target language phrase depends on the previous target words that appear within the phrase. An example showing the influence of word context in phrase translation is given in Table I, based on translation probabilities taken from Chinese→English alignment models described

TABLE I

Bigram Phrase Translation. Likelihood assigned to a phrase translation of a Chinese word is increased when the translation probabilities are conditioned on the English word context.

P(World Trade Center|世贸中心)	=	P(World|世贸中心)		P(Trade|World, 世贸中心)		P(Center|World, Trade, 世贸中心)		
Unigram Approximation	≈	P(World|世贸中心)		P(Trade|世贸中心)		P(Center|世贸中心)		
	=	0.06	×	0.06	×	0.06	=	0.0002
Bigram Approximation	≈	P(World|世贸中心)		P(Trade|World, 世贸中心)		P(Center|Trade, 世贸中心)		
	=	0.06	×	0.99	×	0.99	=	0.0588

later in Section V-B. ‘World Trade Center’ is one of the valid translations for the Chinese word ‘世贸中心’ . The unigram, or context independent, translation of the words ‘World’ , ‘Trade’ , and ‘Center’ yields a much lower likelihood of translation for the entire phrase than when the English word context is taken into account. More complex translation dependencies are clearly worth considering, however any dependencies introduced will necessarily complicate the model; this formulation supports efficient HMM-based estimation and alignment algorithms.

4) Word-to-Phrase Alignment: The idea of explicitly aligning source words to target phrases has been explored for statistical natural language understanding [17], [18]. [17] proposed a statistical word-to-clump generative model with a uniform alignment distribution as is done in IBM Model-1. This “clump” can be thought of as a phrase in the sense used here. Alignment distortions were suggested and studied by [19], and [18] extended the concept of fertility [2] to the generation of phrases, and proposed improved word to phrase translation probabilities by utilizing context. While our model shares this idea of generating multiple target words from a source word, we embed this translation probability as an observation distribution within an HMM alignment process.

5) IBM Models of Word Alignment in Translation: The well-known IBM word alignment models [2] consist of a series of translation alignment models of increasing complexity. The models are generative, in that target words are generated by source words. Model-1 and Model-2 assume the target words are generated independently from the words in the source string, as follows: a source sentence position is selected for each position in the target sentence, and a target word is produced as a translation of the selected source word. In Model-1, the source positions are selected uniformly, while in Model-2 they depend on the target position in question and the lengths of the two strings. Although they are useful, Model-1 and Model-2 are generally considered to be relatively weak models of word translation, mainly because of the extreme simplicity of their alignment and generation processes.

IBM Models 3 and 4 are also generative, although they exchange the source and target directions relative to Models 1 and 2. Models 3 and 4 first decide how many target words each source word should generate; this is specified by the source word *fertility*. For each source sentence position, that many target words are then produced as translations of the source word; there is also a mechanism for generating target words via NULL alignments. The models then arrange the hypothesized target words to produce a target string according to the *distortion models*. In Model-3 target positions are chosen independently for the words generated by each source word. In Model-4, Fig. 3, there are two types of distortion models: one is applied to position the first target word generated by each source word; the second distortion model then determines the relative distance to the previously chosen target word position, i.e. multiple target words might to be inserted into the same position, which would produce an invalid target string. Since Models 3 and 4 assign probability to non-sentences in the target language, they are called *deficient*. Despite this shortcoming, Model-4 word alignments as produced by GIZA++ [7] have been widely used in statistical machine translation systems because of their high quality.

The relationship between our word-to-phrase alignment HMM and the IBM models of fertility and distortion is somewhat complicated. As reviewed, the main features of Model-4 are NULL source words, source word fertility, and the distortion model. The word-to-phrase alignment model also includes NULL source words. However the word-to-phrase alignment HMM describes source word fertility only indirectly. The phrase length distribution does control the number of target words generated by each source word state, but within an alignment each source state can be visited multiple times. Thus the expected number of target words generated by each source word depends on both the phrase length distribution as well as the expected frequency of a source word position in the alignment sequence.

The IBM-4 distortion model, which allows hypothesized words to be distributed throughout the target sentence, is difficult to incorporate into a model that supports efficient dynamic programming search and estimation algorithms. Since algorithmic efficiency is one of our objectives in developing the word-to-phrase alignment model, we avoid this problem by insisting that target words form connected phrases. This is not as general as the Model-4 distortion, although this shortcoming is somewhat balanced by a more powerful (Markov) alignment process. Moreover, the Markov dependencies underlying the word-to-phrase alignment HMM assure that target word subsequences are generated in-place, thereby avoiding deficiency. In summary, the word-to-phrase alignment HMM is not deficient and also supports efficient dynamic programming search and estimation algorithms.

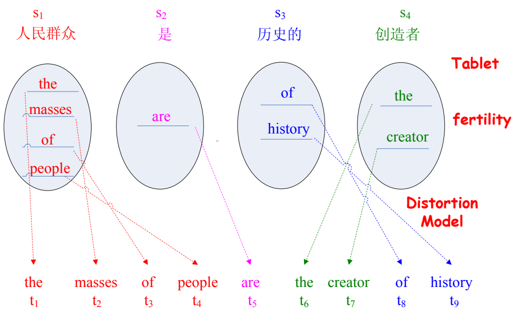

Fig. 3. Simplified Example of Word-to-Phrase Alignments under IBM Model-4.

Despite these differences the word-to-phrase alignment model and Model-4 allow similar word alignments. In the alignment depicted in Fig. 2, for example, Model-4 would allow s_1 to generate t_1, t_3, and t_4 with a fertility of 3, and they would then be distributed as observed under the distortion model. In the word-to-phrase alignment model, since source word positions can generate more than one phrase, s_1 could generate t_1 and t_3t_4 with phrase lengths 1 and 2, respectively. We note that if the alignment process is constrained so that states representing source words are not revisited, the phrase length of the target phrase aligned to the source word is indeed equal to its fertility. However the phrase length component is in general not exactly equivalent to the Model-4 fertility although it has similar descriptive power. It is inspired by features of Model-4, but is incorporated within the word-to-phrase alignment HMM in a manner that allows efficient parameter estimation and alignment procedures.

To summarize these modeling approaches, HMM-based word alignment models and IBM fertility-based models are quite different. The Markov assumption underlying the HMM determines that target words are generated locally, which enables efficient dynamic programming based procedures. The distortion model and fertility information in Model-4 together produce word alignments with better quality than that of HMM but make training procedure computationally complicated. However, both models are based on word-to-word alignment and word-to-word translation. The approach to developing HMM-based word-to-phrase alignment is to make the Markov process more powerful in generating observation sequences: phrases rather than words are emitted after each state transition as shown in Fig. 1. We establish links between source words and target phrases explicitly during the generative procedure, and context within a phrase is also considered in translation. As we will show experimentally later, these features lead to models of comparable performance to Model-4. However the approach remains computationally tractable in both parameter estimation and alignment.

III. Embedded Word-to-Phrase Alignment Model Parameter Estimation

We now discuss estimation of the word-to-phrase alignment model parameters by the EM algorithm. The recursion runs forward, word by word, over the target sentence and gathers statistics relative to the alignment of target phrases to source words.

A. Forward-Backward Procedure

Given a sentence pair (s_1^I, t_1^J), a state space $\{(i,\phi,h) : 1 \le i \le I, 1 \le \phi \le N, h = 0 \text{ or } 1\}$ is created over which the Forward-Backward algorithm will be carried out. The Forward statistic $\alpha_j(i,\phi,h)$ is defined as the probability that the complete source sentence generates the first j words in the target sentence, with the additional constraint that the last ϕ of the target words form a phrase generated by source word s_i . Including the influence of the hallucination process,we introduce the following notation

$$\alpha_j(i,\phi,h) = \begin{cases} P(t_1^j,\ t_{j-\phi+1}^j \leftarrow s_i \mid s_1^I) & h=1 \\ P(t_1^j,\ t_{j-\phi+1}^j \leftarrow \text{NULL} \mid s_1^I) & h=0 \end{cases} .$$

The Forward statistics can be calculated recursively over a trellis of $2 \times N \times I \times J$ nodes as

$$\alpha_j(i,\phi,h) = \{ \sum_{i',\phi',h'} \alpha_{j-\phi}(i',\phi',h') p(i|i',h;I) \} n(\phi; h \cdot s_i)$$

$$\eta \cdot t_1(t_{j-\phi+1}|h \cdot s_i) \prod_{j'=j-\phi+2}^{j} t_2(t_{j'}|t_{j'-1}, h \cdot s_i) \ . \tag{1}$$

The Backward probability $\beta_j(i,\phi,h)$ is defined as the probability that the complete source sentence generates the final $I-j$ target words, given that the target words $t_{j-\phi+1}^j$ form a phrase aligned to $h \cdot s_i$:

$$\beta_j(i,\phi,h) = \begin{cases} P(t_{j+1}^I \mid t_{j-\phi+1}^j \leftarrow s_i\ ,\ s_1^I) & h=1 \\ P(t_{j+1}^I \mid t_{j-\phi+1}^j \leftarrow \text{NULL}\ ,\ s_1^I) & h=0 \end{cases} .$$

It can be calculated recursively over the same trellis as

$$\beta_j(i,\phi,h) = \sum_{i',\phi',h'} \beta_{j+\phi'}(i',\phi',h') p(i'|i,h';I) n(\phi'; h' \cdot s_{i'})$$

$$\eta \cdot t_1(t_{j+1}|h' \cdot s_{i'}) \prod_{j'=j+2}^{j+\phi'} t_2(t_{j'}|t_{j'-1}, h' \cdot s_{i'}) \ . \tag{2}$$

B. Word-to-Phrase Translation Statistics

After the Forward recursion, the conditional probability of sentence $\mathbf{t}$ given $\mathbf{s}$ can be found as

$$P(\mathbf{t} \,|\, \mathbf{s}) = \sum_{i',h',\phi'} P(t_1^J,\ t_{J-\phi'+1}^J \leftarrow h' \cdot s_{i'} \mid \mathbf{s}) \ . \tag{3}$$

The corresponding relationship holds for the Backward probability, as usual.

The probability that a phrase $t_{j-\phi+1}^{j}$ is generated by any of the words in the source sentence can be found as

$$\begin{aligned} & P(\mathbf{t}, t_{j-\phi+1}^{j} \leftarrow h \cdot s_i | \mathbf{s}) \\ = \;& P(t_{j+1}^{I} | t_{j-\phi+1}^{j} \leftarrow h \cdot s_i, s_1^I) \cdot P(t_1^j, t_{j-\phi+1}^{j} \leftarrow h \cdot s_i | s_1^I) \\ = \;& \alpha_j(i, \phi, h) \; \beta_j(i, \phi, h) \end{aligned}$$

With these quantities computed, we can calculate the posterior of any target phrase generated by any source word. Let $\gamma_j(i, \phi, h)$ be the posterior probability that target words $t_{j-\phi+1}^{j}$ form a phrase aligned to the source word $h \cdot s_i$, it can be found as

$$\gamma_j(i, \phi, h) = P(t_{j-\phi+1}^{j} \leftarrow h \cdot s_i | \mathbf{s}, \mathbf{t}) = \frac{\alpha_j(i, \phi, h) \beta_j(i, \phi, h)}{\sum_{i', h', \phi'} \alpha_J(i', \phi', h')}.$$

Finally, reestimation of the Markov transition matrix requires the posterior probability of observing *pairs* of target phrases. The probability that a phrase $t_{j-\phi'+1}^{j}$ and its successor $t_{j+1}^{j+\phi}$ are generated by $h' \cdot s_{i'}$ and $h \cdot s_i$, respectively, can be found as

$$\begin{aligned} P(\mathbf{t}, t_{j-\phi'+1}^{j} \leftarrow h' \cdot s_{i'}, t_{j+1}^{j+\phi} \leftarrow h \cdot s_i | \mathbf{s}) = \alpha_j(i', \phi', h') \eta \\ p(i | i', h; I) n(\phi; h \cdot s_i) p(t_{j+1}^{j+\phi} | s_i, h, \phi) \beta_{j+\phi}(i, \phi, h). \end{aligned} \tag{4}$$

The posterior probability can be found as the ratio of Equation 4 to Equation 3 :

$$\gamma_j(i', \phi', h', i, \phi, h) = P(t_{j-\phi'+1}^{j} \leftarrow h' \cdot s_{i'}, t_{j+1}^{j+\phi} \leftarrow h \cdot s_i \, | \mathbf{t}, \mathbf{s}).$$

C. Parameter Update Relationships

The update equations for the context independent translation table t_1 and the Markov transition probability p_a are given here; the remaining model parameters are updated in a similar manner using statistics collected during the Forward-Backward passes.

Let $\mathbf{T}$ denote the parallel text used as the training set. Let $c(s, t)$ be the posterior counts accumulated over all training sentences of the source word s generating the target word t

$$c(s, t) = \sum_{(\mathbf{s}, \mathbf{t}) \in \mathbf{T}} \sum_{\substack{i, j, \phi, \\ s_i = s}} \gamma_j(i, \phi, h = 1) \; \#_j(t, \phi) \; .$$

Here $\#_j(t, \phi) = \sum_{j'=j-\phi+1}^{j} 1_t(t_{j'})$ is the number of times word t appears in the phrase $t_{j-\phi+1}^{j}$; this is computed directly over the target sentence. The updated estimate of the unigram translation probability is then found as $\hat{t}_1(t|s) = \frac{c(s,t)}{\sum_{t'} c(s,t')}$. The estimation of the bigram translation probability is discussed in Section III-D.2.

The word-to-phrase translation pair statistics are gathered as

$$c(i', i; I) = \sum_{\substack{(\mathbf{s},\mathbf{t}) \in \mathbf{T}, \\ |\mathbf{s}| = I}} \; \sum_{j,\phi',h',\phi} \gamma_j(i', \phi', h', i, \phi, h = 1),$$

where $|\mathbf{s}|$ is the number of words in $\mathbf{s}$. The reestimated transition probability is then computed as

$$\hat{p}_a(i|i'; I) = \frac{c(i', i; I)}{\sum_{i''} c(i'', i; I)}. \tag{5}$$

D. Iterative Estimation Procedures

As with training the IBM fertility-based models [2], [7], the word-to-phrase alignment model parameters are estimated incrementally so that model complexity increases only as training progresses. Any number of training scenarios for the word-to-phrase alignment HMM are possible, however the experiments that will be reported later in this paper were based on the following recipe.

Model parameters are trained from a flat-start without use of any prior alignment information. The final model complexity is determined by the maximum phrase length, N_{max}, which is decided upon beforehand and then verified subsequently through testing of the models.

- Translation and transition tables are initialized as uniform distributions.
- Model-1 parameters are estimated with 10 iterations of EM.
- Model-2 parameters are estimated with 5 iterations of EM.
- The parameters of a word-to-word HMM alignment model are initialized by word alignment counts from Model-2 Viterbi alignments of the parallel text.
- Word-to-word alignment HMM parameters are estimated with 5 iterations of EM.
- For $N = 2, \ldots, N_{max}$:
  - Word-to-phrase HMM parameters are estimated with 5 iterations of EM.
- If Bigram translation tables are to be estimated:
  - Bigram translation tables t_2 are cloned from unigram tables t_1 (at $N = N_{max}$).
  - Word-to-phrase HMMs with bigram-translation tables are estimated with 5 iterations of EM.

This strategy of gradually increasing model complexity as training progresses is motivated by experience in estimating the parameters of large language processing systems, notably the ‘incremental build’ approach to building mixture of Gaussian distribution models in automatic speech recognition [20].

The component distributions that make up the word-to-phrase alignment HMM come together form an extremely complex system. Even with large amounts of parallel text used in training, there is significant risk of overtraining unless preventative steps are taken. We now discuss simple parameter smoothing techniques for robust estimation of the word-to-phrase alignment HMM transition matrices and the bigram translation probabilities.

1) Robust Estimation of Markov Transition Probabilities: When estimated in the usual way (via Equation 5), the transition probabilities $P_a(i|i';I)$ are based on statistics of the conditional expectation that consecutive target phrases are generated by source words in positions i' to i within source sentences of length I. This level of modeling specificity can easily suffer from observation sparsity within the available parallel text, in that the text is partitioned by the source language sentence lengths in estimating the length-specific transition probabilities.

To address this particular problem, Vogel et al.[9] suggested the use of 'jump dependent' Markov transition probabilities, which we adopt here in modified form. The jump transition probability $p_a^{(jump)}(i|i';I)$ is a function only of the 'jump' $i-i'$ made in the alignment sequence. In estimating $p_a^{(jump)}(i|i';I)$, all accumulators corresponding to state transitions with a jump of $i-i'$ contribute to estimating the jump transition probability. The goal is to improve robustness by sacrificing some of the descriptive power of this component so that there are few parameters to estimate.

We employ a simple interpolation scheme to obtain transition probabilities $\hat{p}_a(i|i';I)$ after each iteration of EM. We perform a linear interpolation of the Maximum Likelihood estimate $p_a(i|i';I)$, the 'jump' transition probabilities $p_a^{(jump)}(i|i';I)$ and the uniform distribution $1/I$

$$\tilde{p}_a(i|i';I) = \lambda_1 \cdot \hat{p}_a(i|i';I) + \lambda_2 \cdot p_a^{(jump)}(i|i';I) + \lambda_3 \cdot \frac{1}{I} \tag{6}$$

with $\hat{p}_a$ estimated by the unsmothed EM estimate of Equation 5. The interpolation parameters λ_1, λ_2, and λ_3 are positive, sum to 1, and are tuned over held-out development data.

Performing parameter interpolation in this way does improve robustness, but it is less effective than estimation strategies that control the overall model complexity relative to the amount of relevant training data. We next investigate the use of such techniques for estimation of the bigram translation probabilities.

2) Robust Estimation of Bigram Translation Probabilities: The bigram translation probability assigns likelihood to a target word t which follows another target word t' in a phrase generated as a translation of a given source word s. This probability has the form of a predictive bigram language, $t_2(t|t',s)$, and we borrow techniques from statistical language modeling for its robust estimation. Any of the many backoff schemes for n-gram language modeling could be used, and here we investigate Witten-Bell smoothing [21].

Let $k(t', t, s)$ be the expected number of occurrences of t given that t follows t' in a phrase translated from source word s. Following the notation in Section III-C, the expected number of occurrences are accumulated over all possible relevant word to phrase alignments weighted by their posteriors:

$$k(t', t, s) = \sum_{(\mathbf{s},\mathbf{t}) \in \mathbf{T}} \sum_{\substack{i, j, \phi, \\ s_i = s}} \gamma_j(i, \phi, h = 1) \; \#_{t',t}(t_{j-\phi+1}^{j}) \; ,$$

where $\#_{t',t}(t_{j-\phi+1}^{j})$ is the number of occurrences of bigram $t't$ in the target phrase $t_{j-\phi+1}^{j}$. We choose a threshold L such that the conditional bigram is treated as an unseen event if $k(t', t, s) < L$. For those less frequent events, we back off to the word-to-word translation probabilities, t_1. The total count of seen events is defined as $N(t', s) = \sum_{t:k(t',t,s) \geq L} k(t', t, s)$ and the total seen event types as $T(t', s) = \sum_{t:k(t',t,s) \geq L} 1$. Using these quantities we define

$$\lambda_{t',s} = \frac{T(t', s)}{T(t', s) + N(t', s)}$$

as the total probability mass to be assigned to all unseen events. This probability mass is distributed according to the "unigram" distribution, which is a word-to-word translation probability.

$$t_2(t|t', s) \;\; = \;\; \begin{cases} (1 - \lambda_{t',s}) \frac{k(t',t,s)}{N(t',s)} & k(t', t, s) \geq L \\ \lambda_{t',s} \frac{t_1(t|s)}{r_{t',s}} & \text{otherwise} \end{cases} \qquad (7)$$

where $r_{t',s} = \sum_{t:k(t',t,s) < L} t_1(t|s)$ is introduced for normalization.

IV. Phrase Pair Extraction From Parallel Text

Many current approaches to phrase-based machine translation rely on a phrase translation table, also known as a phrase pair inventory (PPI), which can be extracted from word-aligned parallel text [22]. The PPI is the phrase-level analogue of the word-to-word translation tables that appear in the word-to-phrase alignment HMM, and it forms the basis of phrase-based statistical translation systems.

There have been a variety of schemes proposed to obtain the phrase pairs needed in translation. Bilingual phrase pairs can be induced from word aligned parallel text [23] [22], or indirectly obtained from phrase alignment models estimated over translations [4] [5], and they also can be found by data mining algorithms from parallel strings through co-occurrence analyses [24].

In the widely-used *phrase-extract* algorithm [25], phrase pairs are extracted from word-aligned translation using alignments generated using IBM Model-4 as implemented in the GIZA++ toolkit [7]; the procedure is general and can be applied to sets of alignments generated using other alignment procedures.

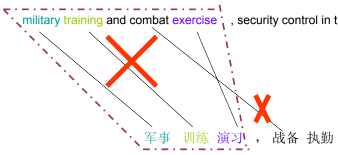
military training and combat exercise , security control in the preparedness against the war
军事 训练 演习 ， 战备 执勤

Fig. 4. Incorrect word-to-word links preventing correct phrase pairs from being found in sentence translations.

For each pair of sentences $\mathbf{s}$ and $\mathbf{t}$, it is assumed that there is a set of alignment links $A = \{(i, j)\}$ that specify how their words are aligned: if $(i, j) \in A$ then $s_i \leftrightarrow t_j$. As a practical matter, this set of alignments has no translation direction. What is usually done, for instance when using IBM Model-4 as implemented in the GIZA++ toolkit [7], is to generate two sets of alignments using one set of models trained in the source-to-target ($\mathbf{s} \rightarrow \mathbf{t}$) direction and another set of models trained in the target-to-source ($\mathbf{s} \leftarrow \mathbf{t}$) direction. The alignment links generated in each translation direction are then merged, and the alignment direction is not retained in forming the merged set of alignments A.

The *phrase-extract* algorithm finds phrase pairs by imposing the following constraint: words within a phrase pair must not align to words outside the phrase pair. Every source phrase and target phrase that satisfy this constraint with respect to the alignment set form a phrase pair.

More formally, let $A = \{(i, j)\}$ be the word alignment link set for the pair of sentences $\mathbf{s}$ and $\mathbf{t}$. Consider a source phrase $s_{i_1}^{i_2} = s_{i_1} \ldots s_{i_2}$ defined by the indices i_1 and i_2, and a target phrase $t_{j_1}^{j_2} = t_{j_1} \ldots t_{j_2}$. The two phrases form a phrase pair $(s_{i_1}^{i_2}, t_{j_1}^{j_2})$ if for every link $(i, j) \in A$, $i_1 \leq i \leq i_2$ iif $j_1 \leq j \leq j_2$. Intuitively, phrase pairs are formed whenever links to words in the phrases respect the phrase boundaries.

We refer to the PPI constructed in this way as the ‘Viterbi Phrase-Extract’ PPI (VPE PPI) since it is derived from Viterbi alignments generated in each translation direction. Once the PPI is determined, the phrase translation probabilities are calculated based on the number of times each phrase pair is extracted from a sentence pair.

The VPE PPI is limited to phrase pairs which can be found in the word alignment set. This can be limiting, as shown in the example of Fig. 4, where the 5^{th} word in the Chinese sentence is incorrectly aligned to the 4^{th} English word. As a consequence the correct phrase pair, depicted by the dotted line, is not identified by the *phrase-extract* algorithm. Various procedures have been proposed to avoid problems such as this (e.g. [6]). We now describe an approach based on posterior distributions defined over phrase pairs as computed under the word-to-phrase alignment HMM.

A. Phrase Pair Induction via Model-Based Posteriors

In considering whether the target phrase $t_{j_1}^{j_2}$ might form a phrase pair with the source phrase $s_{i_1}^{i_2}$ under the *phrase-extract* criterion, we first restate the problem. Given the phrase boundaries i_1, i_2, j_1, j_2, we can define the following set of alignments

$$A(i_1, i_2; j_1, j_2) = \Big\{ \mathbf{a} = a_1^J : a_j \in [i_1, i_2] \text{ iif } j \in [j_1, j_2] \Big\}. \tag{8}$$

For a set of alignments A, the *phrase-extract* criterion can then be stated equivalently as: $(s_{i_1}^{i_2}, t_{j_1}^{j_2})$ form a phrase pair if and only if $A \in A(i_1, i_2; j_1, j_2)$. Defined in this way, there are many possible valid alignments under which $(s_{i_1}^{i_2}, t_{j_1}^{j_2})$ might form a phrase pair, and it is therefore natural to consider the probability of all the alignments in which which this event occurs. The likelihood of the target phrase aligned to the source phrase is obtained by considering all "valid" alignments :

$$P(\mathbf{t}, A(i_1, i_2; j_1, j_2)|\mathbf{s}; \theta) = \sum_{\mathbf{a} \in A(i_1, i_2; j_1, j_2)} P(\mathbf{t}, \mathbf{a}|\mathbf{s}; \theta) \tag{9}$$

Applying Bayes rule, we obtain $P(A(i_1, i_2; j_1, j_2)|\mathbf{s}, \mathbf{t}; \theta) = P(\mathbf{t}, A(i_1, i_2; j_1, j_2)|\mathbf{s}; \theta)/P(\mathbf{t}|\mathbf{s}; \theta)$ which is the posterior probability that $(s_{i_1}^{i_2}, t_{j_1}^{j_2})$ form a phrase pair given $\mathbf{s}$ and $\mathbf{t}$. With this quantity we can consider alternative extraction strategies, such as generating a sorted list of the most likely phrases pairs that can be extracted from a sentence translation pair. This phrase pair posterior distribution (Equation IV-A) applies to any statistical word alignment model. Whether the computation involved is feasible depends on the underlying model. We note that finding the phrase pair posterior distribution under the IBM fertility-based models such as Model-4 faces the same challenges as arise in parameter estimation. For those models there is no efficient way to calculate the likelihood as defined in Equation 9.

B. Phrase Pair Posterior Calculation Under the Word-to-Phrase Alignment HMM

Equation 9 can be computed efficiently under the word-to-phrase alignment Hidden Markov Model using a modified Forward algorithm in which the recursion respects the word alignment constraint of Equation 8. For the given boundaries i_1, i_2, j_1, j_2, the permissible combinations of (i, j, ϕ) are

$$\begin{aligned} i < i_1 \quad &: \quad j < j_1 \quad \text{or} \quad j_2 < j - \phi + 1 \\ i_1 \le i \le i_2 \quad &: \quad j_1 \le j - \phi + 1 \quad \text{and} \quad j \le j_2 \\ i_2 < i \quad &: \quad j < j_1 \quad \text{or} \quad j_2 < j - \phi + 1 \end{aligned}$$

Over these values $\rho_j(i, \phi, h)$ is computed recursively as

$$\Big[\sum_{i',\phi',h'} \rho_{j-\phi}(i', \phi', h') p(i|i', h; I) \Big] \eta \, n(\phi; h \cdot s_i) \times$$

$$t_1(t_{j-\phi+1}|h \cdot s_i) \prod_{j'=j-\phi+2}^{j} t_2(t_{j'}|t_{j'-1}, h \cdot s_i)$$

For all other combinations of indices $\rho_j(i, \phi, h) = 0$. We then find the desired posterior distribution as

$$P(A(i_1, i_2; j_1, j_2)|\mathbf{t}, \mathbf{s}; \theta) = \frac{\sum_{i,\phi,h} \rho_J(i, \phi, h)}{\sum_{i,\phi,h} \alpha_J(i, \phi, h)} .$$

C. Finding Phrase Pairs Under HMM Posterior Distributions

The first step in building an inventory of phrase pairs to be used in translation is to extract all the foreign language phrases from the foreign language text that is to be translated. We need only worry about finding translations for these foreign language phrases. We propose a procedure that builds upon the Viterbi Phrase-Extract PPI as a baseline system. We first identify all the test set phrases for which the Viterbi Phrase-Extract algorithm failed to find translations, and if these occur in the training set, we then use the model-based posterior distribution to find translations for them. The procedure uses word-to-phrase alignment HMMs trained in both translation directions, as follows.

For each foreign phrase v not in the VPE PPI , find all sentence pairs in which v appears, i.e. find all pairs (t_1^J, s_1^I) and j_1, j_2 such that $t_{j_1}^{j_2} = v$. For each pair and for $i_1, i_2 : \ 1 \leq i_1 \leq i_2 \leq I$, compute

$$b(i_1, i_2) = P_{\mathbf{t} \to \mathbf{s}}(\, A(i_1, i_2; j_1, j_2) \,|\, s_1^I, t_1^J)$$

$$f(i_1, i_2) = P_{\mathbf{s} \to \mathbf{t}}(\, A(i_1, i_2; j_1, j_2) \,|\, s_1^I, t_1^J)$$

$$g(i_1, i_2) = \sqrt{f(i_1, i_2) \, b(i_1, i_2)}$$

find $(\hat{i}_1, \hat{i}_2) = \text{argmax}_{1 \leq i_1, i_2 \leq l} \; g(i_1, i_2)$, and set $u = s_{\hat{i}_1}^{\hat{i}_2}$. Add (u, v) to the PPI if any of the following three conditions hold: $b(\hat{i}_1, \hat{i}_2) \geq T_g$ and $f(\hat{i}_1, \hat{i}_2) \geq T_g$, or $b(\hat{i}_1, \hat{i}_2) < T_g$ and $f(\hat{i}_1, \hat{i}_2) > T_p$, or $f(\hat{i}_1, \hat{i}_2) < T_g$ and $b(\hat{i}_1, \hat{i}_2) > T_p$.

The first condition extracts phrase pairs based on the geometric mean of the $\mathbf{s} \to \mathbf{t}$ and $\mathbf{t} \to \mathbf{s}$ posteriors. The threshold T_p selects additional phrase pairs under a more forgiving criterion: as T_p decreases, more phrase pairs are added and the PPI coverage increases. A balance between coverage and phrase translation quality can be achieved by varying the thresholds. Note that this algorithm is constructed specifically to improve a Viterbi PPI; it is certainly not the only way to extract phrase pairs under the phrase-to-phrase posterior distribution. An alternative approach would omit the Viterbi Phrase-Extract procedure entirely.

V. Experiments in Chinese-English Word Alignment

Automatic word alignment techniques developed for phrase-based statistical machine translation must ultimately be assessed in terms of translation quality. However word alignment quality can also be measured directly by comparison to parallel text which has been aligned by bilingual human annotators. Such intermediate measures of alignment quality have been found to be predictive, if not determinant, of translation quality and can also provide insight into the behavior of the modeling procedures which generated the alignments.

We report results of Chinese-English word alignment experiments using Alignment Error Rate (AER) [7] as the measure of alignment quality. AER measures performance relative to a set of reference word alignments produced over a test set by bilingual human annotators, and is defined as

$$AER(B;B') = 1 - 2 \times \frac{|B \cap B'|}{|B'| + |B|}$$

where B is the set of reference word links, and B' are the automatically generated word links; we note that we use only the simplest form of AER defined by [7]. Our experiments will be based on an alignment test set consisting of 124 sentences from the NIST 2001 dry-run [26] which have been manually word aligned. We also define variations of AER over word-to-word and word-to-phrase links. These additional AER measures make it possible to measure alignment quality over these different categories of links as well as the relationship of both to the overall AER. We partition the reference alignments into two sets: the set B_{1-1} contains word-to-word reference links (e.g. $s_1 \rightarrow t_1$ in Fig. 2), and the set B_{1-N} contains word-to-phrase reference links (e.g. $s_1 \rightarrow t_3 t_4$ in Fig. 2). The automatically generated word alignments B' are partitioned similarly. With these partitioned link sets, we define two variants of AER: $AER_{1-1} = AER(B_{1-1}; B'_{1-1})$ and $AER_{1-N} = AER(B_{1-N}; B'_{1-N})$, which measure word-to-word and word-to-phrase alignment quality, respectively. We note that these measures are somewhat harsh as defined. For example, hypothesized word-to-word links without a corresponding link in B_{1-1} are counted as spurious even if the needed link is present within word-to-phrase links in B_{1-N}.

A. Words and Phrases in the Reference Alignments

In the manually generated reference alignments, of the Chinese words which are aligned to more than one English words, 82% of these words align with consecutive English words (phrases). In the other direction, among all English words which are aligned to multiple Chinese words, 88% of these align to Chinese phrases. In this collection, at least, observations of word-to-phrase alignments are plentiful.

B. Alignment Model Estimation Over Parallel Corpora of Intermediate Size

We present word alignment experiments on an FBIS Chinese/English parallel corpus which consists of 11,537 parallel documents with approximately 10M English and 7.5M Chinese words. The Chinese documents are first segmented into words by the LDC word segmenter [27]. The parallel documents are then aligned via an iterative, unsupervised procedure [28] which yields aligned translation segments of approximately sentence length, many of which are sub-sentence fragments. This leads to a reduction in the estimation time for statistical word alignment models, and an increase in the amount of parallel text which can be used in training, since some pairs with overly long sentences would otherwise be discarded as a practical matter [28].

The AER, AER_{1-1}, and AER_{1-N} measures of alignment quality are presented in Table II for word-to-phrase HMM alignments generated by the procedure described in Section III. Measurements of Model 4 alignments generated by GIZA++ are also included for comparison. We first note that, as has been previously reported [7], word-to-word alignment HMMs do not match IBM Model 4 in Alignment Error Rate. However we find that AER can be improved in HMM-based alignment by introducing word-to-phrase translation dependencies. For the word-to-phrase alignment HMM estimated in the Chinese→English direction, we see reduced AER for phrase lengths up to four words ($N = 4$). AER is also reduced in the English→Chinese direction for phrase lengths of two words ($N = 2$), although alignment performance then degrades for phrases of length greater than two (not presented here).

We now analyze the effect of including the bigram phrase translation probability. The intended role of this component is to increase the likelihood of specific word-to-phrase alignments within the overall model. When this component is added to the best word-to-phrase alignment model, we find that the word-to-phrase Alignment Error Rate (AER_{1-N}) does decrease both in the Chinese→English and the English→Chinese direction. However this reduction is associated with an increase in word-to-word Alignment Error Rate (AER_{1-1}) which is due to a drop in recall as fewer word-to-word alignments are produced. For Chinese→English alignment, AER is still reduced overall; however, in English→Chinese alignment, overall AER increases.

This behavior can be explained by looking at the distribution of the word-to-word and word-to-phrase links that are produced under the model. The balance between the number of each can be controlled by the phrase count parameter η : as η increases, the model favors alignments with more word-to-word links and fewer word-to-phrase links. This behavior is exhibited in Table III which tabulates the links generated in the C→E direction under the word-to-phrase alignment HMM of maximum phrase length

TABLE II

CHINESE-ENGLISH WORD ALIGNMENT ERROR RATES. MODELS ARE ESTIMATED OVER THE FBIS COLLECTION. ALIGNMENTS ARE COMPARED TO REFERENCE WORD ALIGNMENTS PRODUCED BY BILINGUAL HUMAN ANNOTATORS.

	Model	AER_{1-1}	AER_{1-N}	AER
Chinese→English	Model-4	37.9	68.3	37.3
	Word-to-Word HMM	42.8	72.9	42.0
	Word-to-Phrase HMM, N=2	38.3	71.2	38.1
	Word-to-Phrase HMM, N=3	37.4	69.5	37.8
	Word-to-Phrase HMM, N=4	37.1	69.1	37.8
	+ bigram phrase translation	37.5	65.8	37.1
English→Chinese	Model-4	42.3	87.2	45.0
	Word-to-Word HMM	45.0	90.6	47.2
	Word-to-Phrase HMM, N=2	42.7	87.5	44.5
	+ bigram phrase translation	44.2	85.5	45.1

$N = 4$ with bigram phrase translation. The proportion of word-to-phrase links increases with η, and the overall Alignment Error Rate suggests a good balance at $\eta = 8.0$.

In these experiments, the following configuration achieves comparable Alignment Error Rate to IBM Model 4 alignment: for $\mathbf{s} \rightarrow \mathbf{t}$, we set $N = 4$, $\eta = 8.0$, and include the bigram translation component; for $\mathbf{t} \rightarrow \mathbf{s}$, we set $N = 2$, $\eta = 8.0$, and leave out the bigram translation component. The parameters $\lambda_1, \lambda_2, \lambda_3$ are set to 0.36, 0.4, and 0.24, in both translation directions. We will use these settings for all subsequent experiments.

TABLE III

WORD-TO-WORD AND WORD-TO-PHRASE LINKS AS A FUNCTION OF THE PHRASE COUNT PARAMETER η.

η	Word-to-Word Links	Word-to-Phrase Links	Total Links	AER
2	1677	1948	3625	40.7
4	1966	1550	3516	38.4
6	2140	1310	3450	38.6
8	2252	1146	3398	38.1
10	2368	1010	3378	39.2
12	2448	921	3369	39.4

C. Alignment Model Estimation Over Large Parallel Corpora

The results reported in the previous section are not the first published experiments in which novel alignment techniques have proven to be competitive with Model 4. However many of these results were obtained on small to medium sized collections of parallel text. On larger collections of parallel text, Model 4 consistently produces alignments of higher quality than those of novel alternatives (e.g. [7], [10]).

The word-to-phrase alignment HMMs have been developed with the specific intent of applying them to the large parallel text collections used in statistical machine translation. We therefore wish to verify that the performance of these models remains comparable to that of Model 4 as the parallel text increases in size. We investigate alignment performance with models trained over the several collections of Chinese-English parallel text available from LDC as of June 2005. The "NEWS" collection refers to the LDC parallel Chinese/English news corpora (mainly FBIS, Xinhua, Hong Kong News, Hong Kong Hansard, Sinorama, and the Chinese Treebank). The "NEWS+UN01-02" collection consists of the "NEWS" collections along with the United Nations Chinese-English parallel documents from the years 2001 and 2002. Finally, the "ALL C-E" collection consists of all the C-E parallel text available from LDC as of June 2005. This consists of the NEWS corpora with the UN translations from 1993 through 2002.

In the experiments reported in Table IV, we find that word-to-phrase alignment HMM performance is comparable to that of Model-4 over all these collections. We do note a small degradation in the English→Chinese alignments under the word-to-phrase alignment HMM. It is quite possible that this one-to-many model suffers slightly with English as the source language and Chinese as the target language, since English sentences tend to be longer. Notably, simply increasing the amount of parallel text used in training need not improve AER. However, larger aligned collections can give improved phrase pair coverage of the test set and can lead to improved translation performance.

1) Efficient Parameter Estimation Over Large Parallel Text Collections: One of the desirable aspects of estimation of HMM parameters is that the Forward-Backward steps can be run in parallel. In this application, the parallel text is partitioned into subsets of sentence pairs; the Forward-Backward algorithm is run over the subsets on different CPUs; and statistics are merged to re-estimate model parameters. Partitioning the parallel text can also reduce the memory usage of individual Forward-Backward steps, since different co-occurrence tables can be kept for each partition. With the "ALL C-E" parallel text collection, a single set of word-to-phrase alignment HMMs can be trained over 200M words of Chinese-English parallel text by splitting the parallel text into 40 subsets: each Forward-Backward process takes

TABLE IV

CHINESE-ENGLISH WORD ALIGNMENT PERFORMANCE OVER LARGE PARALLEL TEXT COLLECTIONS. ALIGNMENTS ARE GENERATED IN THE CHINESE→ENGLISH AND ENGLISH→CHINESE DIRECTIONS. ALIGNMENT ERROR RATE (AER) IS MEASURED OVER A WORD-ALIGNED 124 SENTENCE SUBSET OF THE CHINESE-ENGLISH FBIS PARALLEL TEXT CORPUS.

Parallel Text	English Words	Model	$AER_{C\to E}$	$AER_{E\to C}$
NEWS	71M	Model 4	37.1	45.3
		Word-to-Phrase HMM	36.1	44.8
NEWS + UN01-02	96M	Model 4	36.1	43.4
		Word-to-Phrase HMM	36.4	44.2
ALL C-E	200M	Word-to-Phrase HMM	36.8	44.7

less than 2GB of memory and the training procedure finished in five days. By contrast, the 96M English word "NEWS+UN01-02" is about the largest C-E parallel text over which we can train Model-4 with our GIZA++ configuration and computing infrastructure.

VI. ARABIC-ENGLISH AND CHINESE-ENGLISH STATISTICAL MACHINE TRANSLATION

We now report automatic translation performance in Arabic to English and Chinese to English translation. The statistical machine translation systems are based on statistics extracted from parallel text by word-to-phrase HMM alignment and by the GIZA++ implementation of IBM Model-4, for comparison. Automatic translation is performed using the Transducer Translation Model (TTM) [29], [30].

The model generates a source language sentence under a source language model, which in practice is a standard back-off n-gram language model. The source sentence is then segmented into sequences of source phrases, producing a lattice of source language phrase sequences. These source phrases are translated into phrases in the target language. A phrase reordering component [30] is applied to map the target phrases from source phrase order into target phrase order. Target phrases are allowed to be inserted with a phrase-length dependent probability; when the generative model is used in translation and the target language sequence is fixed, this step permits foreign phrases to be deleted rather than translated. The target phrase sequences are finally mapped to target sentences. The phrase reordering component is a finite state systems that allows adjacent phrases to be reversed with a phrase-pair dependent reordering probability. The model reported here allows for a maximum jump of 1, i.e. phrases are reordered only with their immediate neighbors, and the probability or reordering is found by an embedded parameter reestimation over the parallel text. The reestimation procedure is carried out as a form of Viterbi training, hence the reordering component is referred to as MJ-1 VT. Some translation results are reported with the reordering step omitted, a configuration referred to as monotone phrase order translation. Although this configuration is suboptimal, it is fast and the results produced are useful in system development.

We will report translation performance of two Chinese→English translation systems. The smaller system is based on statistics extracted from the FBIS C-E parallel text collection. The language model used for this system is a word trigram language model estimated using 21M words taken from the English side of the parallel text; all language models in this article are built with the SRILM toolkit using modified Kneser-Ney smoothing [31]. The larger system is based on alignments generated over all available C-E parallel text (the "ALL C-E" collection of Section V-C). The language model is an equal-weight interpolated trigram model trained over 373M English words taken from the English side of the parallel text and the LDC English Gigaword corpus.

We make use of a large and a small Arabic→English translation system. In the small system the training material is from the A-E News parallel corpus, with $\sim$3.5M words on the English side and the Arabic text tokenized by the Buckwalter analyzer [32]. The language model is an equal-weight interpolated trigram built over $\sim$400M words from the English side of the collection, including the UN collections, and the LDC English Gigaword collection. The large Arabic/English system employs the same language model, but is based on statistics extracted from all Arabic-English parallel text available from the LDC as of June 2005; this consists of approximately 130M words on the English side.

A. *Word-to-Phrase HMM Alignment and IBM Model-4 Alignment*

We report performance on the NIST Chinese/English 2002, 2003, and 2004 (News only) MT evaluation sets, and the NIST Arabic/English 2002, 2003, and 2004 (News only) MT evaluation sets. These consist of 878, 919, and 901 Chinese sentences and 1043, 663, and 707 Arabic sentences, respectively, along with four English reference translations for each. Translation performance is measured through the BLEU [33] metric relative to the four reference translations.

We first look at translation performance of A→E and C→E systems based on alignment models trained over the small collections of parallel text. In estimating the word-to-phrase alignment HMMs, we follow the alignment procedure and model configuration specified in Section V-B. The baseline systems (Table VI and V, line 1) are based on Model-4 Viterbi Phrase-Extract PPIs, also described in Section V-B.

We compare word-to-phrase HMM alignments directly to Model-4 alignments by using the Viterbi Phrase-Extract (V-PE) procedure to extract phrase translation tables from the word-to-phrase HMM alignments. In C→E translation (Table VI, line 3), performance is comparable to that of Model-4, whereas

TABLE V

Arabic→English Monotone Translation and PPI Extraction Procedures

Parallel		V-PE		eval02		eval03		eval04	
Text		Model	T_p	cvg	BLEU	cvg	BLEU	cvg	BLEU
A-E News	1	Model 4	-	19.5	36.9	21.5	39.1	18.5	40.0
	2		0.7	23.8	37.6	26.6	40.2	22.4	40.3
	3	Word-to-Phrase HMM	-	18.4	36.2	20.6	38.6	17.4	39.2
	4		1.0	21.8	36.7	24.3	39.3	20.4	39.7
	5		0.9	23.2	37.2	25.8	39.7	21.8	40.1
	6		0.7	23.7	37.2	26.5	39.7	22.4	39.9
	7		0.5	24.0	37.2	26.9	39.7	22.7	39.8
All A-E	8	Model 4	-	26.4	38.1	28.1	40.1	28.2	39.9
	9	Word-to-Phrase HMM	-	24.8	38.1	26.6	40.1	26.7	40.6
	10		0.7	30.7	39.3	32.9	41.6	32.5	41.9

in A→E translation (Table V, line 3), performance lags slightly.

We next investigate the Phrase-Posterior Augmentation procedure (Table VI, Table V, lines 4-7). We begin with the PPI based on phrase pairs extracted from the word-to-phrase HMM alignments by the Viterbi Phrase-Extract procedure. This inventory is increased by adding phrase pairs based on their phrase-posterior probability; as the cut-off threshold T_p decreases, more phrase pairs will be added. By augmenting the PPI in this way we obtain a $\sim 1\%$ improvement in BLEU score; the value of $T_p = 0.7$ gives improvements in both translation systems across all evaluation sets. In C→E translation, this yields good gains relative to Model-4, while in A→E we match or improve the Model-4 performance.

The performance gains through PPI augmentation are consistent with increased PPI coverage of the test set (the 'cvg' values in Table VI and Table V), where coverage is measured as the percentage of test set phrases that appear in each of the PPIs. Roughly speaking, if the coverage was 100%, the PPI would contain at least one translation for every target phrase to be translated. The augmentation scheme is designed specifically to increase coverage and we find that BLEU score improvements do indeed track the phrase coverage of the test set. This is further confirmed by the experiment of Table VI and Table V, line 2 in which we take the PPI extracted from Model-4 Viterbi alignments and add phrase pairs to it using the Phrase-Posterior augmentation scheme with $T_p = 0.7$. We find that the augmentation scheme under the word-to-phrase alignment HMMs can be used to improve the Model-4 PPI itself.

We also investigate C→E and A→E translation performance with PPIs extracted from large parallel

TABLE VI

CHINESE→ENGLISH MONOTONE PHRASE TRANSLATION AND PPI EXTRACTION PROCEDURES

Parallel		V-PE		eval02		eval03		eval04	
Text		Model	T_p	cvg	BLEU	cvg	BLEU	cvg	BLEU
FBIS	1	Model 4	-	20.1	23.8	17.7	22.8	20.2	23.0
	2		0.7	24.6	24.6	21.4	23.7	24.6	23.7
	3	Word-to-Phrase HMM	-	19.7	23.9	17.4	23.3	19.8	23.3
	4		1.0	23.1	24.0	20.0	23.7	23.2	23.5
	5		0.9	24.0	24.8	20.9	23.9	24.0	23.8
	6		0.7	24.6	24.9	21.3	24.0	24.7	23.9
	7		0.5	24.9	24.9	21.6	24.1	24.8	23.9
All C-E	8	Model 4	-	32.5	27.7	29.3	27.1	32.5	26.6
	9	Word-to-Phrase HMM	-	30.6	27.9	27.5	27.0	30.6	26.4
	10		0.7	38.2	28.2	32.3	27.3	37.1	26.8

text collections. Performance of systems based on Model-4 Viterbi Phrase-Extract PPIs is shown in Table VI and Table V, line 8; these parameter values were fixed based on the values determined in the experiments over the smaller parallel texts. To train Model-4 using GIZA++, we split the parallel text into two (A-E) or three (C-E) partitions, and train models for each division separately, and find word alignments for each division separately with their models; we find that memory usage is otherwise too great. These serve as a single set of alignments for the parallel text, as if they had been generated under a single alignment model. When we translate with Viterbi Phrase-Extract PPIs taken from Word-to-Phrase HMM alignments created over all available parallel text, we find comparable performance to the Model-4 baseline (Table VI, Table V, line 9). Using the Phrase-Posterior augmentation scheme with $T_p = 0.7$ yields further improvement (Table VI, Table V, line 10).

TABLE VII

MONOTONE TRANSLATION ON MERGED ARABIC→ENGLISH AND CHINESE→ENGLISH TEST SETS.

Model	PPI	$BLEU_{C-E}$	$BLEU_{A-E}$
Model-4	baseline	$27.29^{\pm 0.5}$	$39.39^{\pm 0.6}$
Word-to-Phrase HMM	augmented	$27.47^{\pm 0.5}$	$40.48^{\pm 0.6}$

We also perform tests to see if the improvements under the BLEU metric are statistically significant

[34]. Pooling all three test sets of eval02, eval03, and eval04, we form large test sets for C→E and A→E translations. We compare translation performance of two setups: one is the Model-4 word alignments with the baseline PPI (as in Table VI and V, line 8), the other is the word-to-phrase word alignments with the augmented PPI (as in Table VI and V, line 10). We show their BLEU scores as well as their 95% confidence intervals in Table VII. We find that the Word-to-Phrase alignment model leads to equivalent C→E system performance as that of Model-4, while A→E system improvements are significant at a 95% level [34].

The final series of results are reported in Table VIII in which we show that the word-to-phrase HMM alignments and the phrase pair augmentation procedure can be used with more complex translation systems based on phrase reordering models and higher order ngram English language models.

TABLE VIII

ARABIC→ENGLISH AND CHINESE→ENGLISH TRANSLATION PERFORMANCE WITH 3GRAM AND 4GRAM ENGLISH LANGUAGE MODELS AND MJ-1 VT LOCAL PHRASE REORDERING MODELS.

Parallel	V-PE		Phrase	NGram	eval02	eval03	eval04
Text	Model	T_p	Order	Order	BLEU	BLEU	BLEU
All A-E	Model 4	-	monotone	3	38.2	40.1	
	Word-to-Phrase HMM	0.7	monotone	3	39.3	41.6	41.9
		-	MJ-1 VT	3	41.5	43.4	
		0.7	MJ-1 VT	3	42.3	43.9	45.1
		0.7	MJ-1 VT	4	43.1	45.0	45.6
All C-E	Model 4	-	monotone	4	28.5	27.4	
	Word-to-Phrase HMM	0.7	monotone	4	28.9	27.4	27.3
		0.7	MJ-1 VT	4	30.2	28.2	28.9

VII. CONCLUSION AND FUTURE WORK

We have described an alignment methodology for statistical machine translation that is motivated by features of IBM Model-4 but based on computationally efficient hidden Markov models. We have developed alternative formulations of fertility and distortion and combined them with refinements such as bigram translation probabilities and have done so while maintaining computational tractability and avoiding model deficiency. These components have been analyzed through various alignment experiments and their role in capturing various alignment phenomena has been discussed. Effective training and alignment procedures have been developed, and in Arabic-English and Chinese-English translation the word-to-phrase alignment HMM can be used to generate alignments of comparable quality to those of IBM Model-4 even over large collections of parallel text.

We have also shown that exact computation of posterior statistics under these models can be used to augment phrase-pairs extracted from word aligned parallel text. This leads to improved translation performance, but it also demonstrates the power of the modeling approach. Although the model topology explicitly describes only word-to-phrase alignments, through marginalization we can describe more complex phenomena, such as phrase-to-phrase alignment. This is done by introducing the quantity of interest - the set of alignments consistent with phrase-to-phrase alignments - and computing posterior distributions over such sets. The key is to formulate the desired sets so that the computation can be carried out efficiently with respect to the model topology. In closing we note that these models are still relatively simple and we expect further improvements in alignment and translation quality as we refine them.

Acknowledgments We would like to thank Shankar Kumar for providing the TTM translation system.

References

[1] W. A. Gale and K. W. Church, "A program for aligning sentences in bilingual corpora," in *Meeting of the Association for Computational Linguistics*, 1991, pp. 177–184.

[2] P. Brown, S. D. Pietra, V. D. Pietra, and R. Mercer, "The mathematics of machine translation: Parameter estimation," *Computational Linguistics*, vol. 19, pp. 263–312, 1993.

[3] I. D. Melamed, "Models of translational equivalence among words," *Comput. Linguist.*, vol. 26, no. 2, pp. 221–249, 2000.

[4] D. Marcu and W. Wong, "A phrase-based joint probability model for statistical machine translation," in *Proc. of EMNLP*, 2002.

[5] Y. Zhang and S. Vogel, "An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora," in *Proceedings of the Tenth Conference of the European Association for Machine Translation*, 2005.

[6] A. Venugopal, S. Vogel, and A. Waibel, "Effective phrase translation extraction from alignment models," in *Proc. of ACL*, 2003.

[7] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[8] F. Och and H. Ney, "Improved statistical alignment models," in *Proc. of ACL*, Hong Kong, China, 2000.

[9] S. Vogel, H. Ney, and C. Tillmann, "HMM based word alignment in statistical translation," in *Proc. of COLING*, 1996.

[10] K. Toutanova, H. T. Ilhan, and C. Manning, "Extentions to HMM-based statistical word alignment models," in *Proc. of EMNLP*, 2002.

[11] F. Jelinek, *Statistical methods for speech recognition.* Cambridge, MA, USA: MIT Press, 1997.

[12] Y. Deng and W. Byrne, "Hmm word and phrase alignment for statistical machine translation," in *Proc. of HLT-EMNLP*, 2005.

[13] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMMs to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 4, pp. 360–378, 1996.

[14] S. E. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer, Speech, and Language*, vol. 1, no. 1, pp. 29–45, 1986.

[15] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "Word-sense disambiguation using statistical methods," in *Proc. of ACL*, 1991, pp. 264–270.

[16] A. L. Berger, S. D. Pietra, and V. J. D. Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[17] M. Epstein, K. Papineni, S. Roukos, T. Ward, and S. D. Pietra, "Statistical natural language understanding using hidden clumpings," in *Proceedings of ICASSP*, vol. 1, Atlanta, GA, May 1996, pp. 176–179.

[18] S. D. Pietra, M. Epstein, S. Roukos, and T. Ward, "Fertility models for statistical natural language understanding," in *Proc. of EACL*, 1997, pp. 168–173.

[19] M. Epstein, "Statistical source channel models for natural language understanding," Ph.D. dissertation, New York University, September 1996.

[20] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book, Version 3.1*, Dec. 2001.

[21] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," in *IEEE Trans. Inform Theory*, vol. 37, July 1991, pp. 1085–1094.

[22] P. Koehn, F. Och, and D. Marcu, "Statistical phrase-based translation." in *Proc. of HLT-NAACL*, 2003.

[23] F. Och, C. Tillmann, and H. Ney, "Improved alignment models for statistical machine translation," in *Proc. of the Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, USA, 1999, pp. 20–28.

[24] K. Yamamoto, T. Kudo, Y. Tsuboi, and Y. Matsumoto, "Learning sequence-to-sequence correspondences from parallel corpora via sequential pattern mining," in *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, R. Mihalcea and T. Pedersen, Eds. Edmonton, Alberta, Canada: Association for Computational Linguistics, May 31 2003, pp. 73–80.

[25] F. Och, "Statistical machine translation: From single word models to alignment templates," Ph.D. dissertation, RWTH Aachen, Germany, 2002.

[26] NIST, *The NIST Machine Translation Evaluations*, 2004, http://www.nist.gov/speech/tests/mt/.

[27] LDC, *LDC Chinese Segmenter*, 2002, http://www.ldc.upenn.edu/Projects/Chinese.

[28] Y. Deng, S. Kumar, and W. Byrne, "Segmentation and alignment of parallel text for statistical machine translation," *Journal of Natural Language Engineering*, vol. 12, no. 4, 2006.

[29] S. Kumar, Y. Deng, and W. Byrne, "A weighted finite state transducer translation template model for statistical machine translation," *Journal of Natural Language Engineering*, vol. 12, no. 1, 2006.

[30] S. Kumar and W. Byrne, "Local phrase reordering models for statistical machine translation," in *Proc. of HLT-EMNLP*, 2005.

[31] A. Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. of ICSLP*, 2002.

[32] LDC, *Buckwalter Arabic Morphological Analyzer Version 1.0*, 2002, LDC Catalog Number LDC2002L49.

[33] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a method for automatic evaluation of machine translation," IBM Research Division, Tech. Rep. RC22176 (W0109-022), 2001.

[34] F. Och, "Minimum error rate training in statistical machine translation," in *Proc. of ACL*, 2003.