Gappy Phrasal Alignment by Agreement

Mohit Bansal[*]
UC Berkeley, CS Division
mbansal@cs.berkeley.edu

Chris Quirk
Microsoft Research
chrisq@microsoft.com

Robert C. Moore
Google Research
robert.carter.moore@gmail.com

Abstract

We propose a principled and efficient phrase-to-phrase alignment model, useful in machine translation as well as other related natural language processing problems. In a hidden semi-Markov model, word-to-phrase and phrase-to-word translations are modeled directly by the system. Agreement between two directional models encourages the selection of parsimonious phrasal alignments, avoiding the overfitting commonly encountered in unsupervised training with multi-word units. Expanding the state space to include "gappy phrases" (such as French *ne* $\star$ *pas*) makes the alignment space more symmetric; thus, it allows agreement between discontinuous alignments. The resulting system shows substantial improvements in both alignment quality and translation quality over word-based Hidden Markov Models, while maintaining asymptotically equivalent runtime.

1 Introduction

Word alignment is an important part of statistical machine translation (MT) pipelines. Phrase tables containing pairs of source and target language phrases are extracted from word alignments, forming the core of phrase-based statistical machine translation systems (Koehn et al., 2003). Most syntactic machine translation systems extract synchronous context-free grammars (SCFGs) from aligned syntactic fragments (Galley et al., 2004; Zollmann et al., 2006), which in turn are derived from bilingual word alignments and syntactic parses. Alignment is also used in various other NLP problems such as entailment, paraphrasing, question answering, summarization and spelling correction.

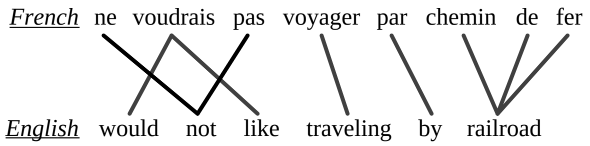

Figure 1: French-English pair with complex word alignment.

A limitation to word-based alignment is undesirable. As seen in the French-English example in Figure 1, many sentence pairs are naturally aligned with multi-word units in both languages (*chemin de fer*; *would* $\star$ *like*, where $\star$ indicates a gap). Much work has addressed this problem: generative models for direct phrasal alignment (Marcu and Wong, 2002), heuristic word-alignment combinations (Koehn et al., 2003; Och and Ney, 2003), models with pseudo-word collocations (Lambert and Banchs, 2006; Ma et al., 2007; Duan et al., 2010), synchronous grammar based approaches (Wu, 1997), etc. Most have a large state-space, using constraints and approximations for efficient inference.

We present a new phrasal alignment model based on the hidden Markov framework (Vogel et al., 1996). Our approach is semi-Markov: each state can generate multiple observations, representing word-to-phrase alignments. We also augment the state space to include contiguous sequences. This corresponds to phrase-to-word and phrase-to-phrase alignments. We generalize alignment by agreement (Liang et al., 2006) to this space, and find that agreement discourages EM from overfitting. Finally, we make the alignment space more symmetric by including *gappy* (or non-contiguous) phrases. This allows agreement to reinforce non-contiguous align-

[*]Author was a summer intern at Microsoft Research during this project.

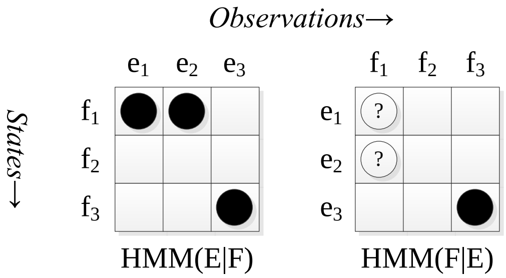

Figure 2: The model of E given F can represent the phrasal alignment $\{e_1, e_2\} \sim \{f_1\}$. However, the model of F given E cannot: the probability mass is distributed between $\{e_1\} \sim \{f_1\}$ and $\{e_2\} \sim \{f_1\}$. Agreement of the forward and backward HMM alignments tends to place less mass on phrasal links and greater mass on word-to-word links.

ments, such English *not* to French *ne* $\star$ *pas*. Pruning the set of allowed phrases preserves the time complexity of the word-to-word HMM alignment model.

1.1 Related Work

Our first major influence is that of *conditional phrase-based models*. An early approach by Deng and Byrne (2005) changed the parameterization of the traditional word-based HMM model, modeling subsequent words from the same state using a bigram model. However, this model changes only the parameterization and not the set of possible alignments. More closely related are the approaches of Daumé III and Marcu (2004) and DeNero et al. (2006), which allow phrase-to-phrase alignments between the source and target domain. As DeNero warns, though, an unconstrained model may overfit using unusual segmentations. Interestingly, the phrase-based hidden semi-Markov model of Andrés-Ferrer and Juan (2009) does not seem to encounter these problems. We suspect two main causes: first, the model interpolates with Model 1 (Brown et al., 1994), which may help prevent overfitting, and second, the model is monotonic, which screens out many possible alignments. Monotonicity is generally undesirable, though: almost all parallel sentences exhibit some reordering phenomena, even when languages are syntactically very similar.

The second major inspiration is *alignment by agreement* by Liang et al. (2006). Here, soft intersection between the forward (F→E) and backward (E→F) alignments during parameter estimation produces better word-to-word correspondences. This unsupervised approach produced alignments with incredibly low error rates on French-English, though only moderate gains in end-to-end machine translation results. Likely this is because the symmetric portion of the HMM space contains only single word to single word links. As shown in Figure 2, in order to retain the phrasal link $f_1 \sim e_1, e_2$ after agreement, we need the reverse phrasal link $e_1, e_2 \backsim f_1$ in the backward direction. However, this is not possible in a word-based HMM where each observation must be generated by a single state. Agreement tends to encourage 1-to-1 alignments with very high precision and but lower recall. As each word alignment acts as a constraint on phrase extraction, the phrase-pairs obtained from those alignments have high recall and low precision.

2 Gappy Phrasal Alignment

Our goal is to unify phrasal alignment and alignment by agreement. We use a phrasal hidden semi-Markov alignment model, but without the monotonicity requirement of Andrés-Ferrer and Juan (2009). Since phrases may be used in both the state and observation space of both sentences, agreement during EM training no longer penalizes phrasal links such as those in Figure 2. Moreover, the benefits of agreement are preserved: meaningful phrasal links that are likely in both directions of alignment will be reinforced, while phrasal links likely in only one direction will be discouraged. This avoids segmentation problems encountered by DeNero et al. (2006).

Non-contiguous sequences of words present an additional challenge. Even a semi-Markov model with phrases can represent the alignment between English *not* and French *ne* $\star$ *pas* in one direction only. To make the model more symmetric, we extend the state space to include *gappy phrases* as well.[1] The set of alignments in each model becomes symmetric, though the two directions model gappy phrases differently. Consider *not* and *ne* $\star$ *pas*: when predicting French given English, the alignment corresponds to generating multiple distinct ob-

[1] We only allow a single gap with one word on each end. This is sufficient for the vast majority of the gapped phenomena that we have seen in our training data.

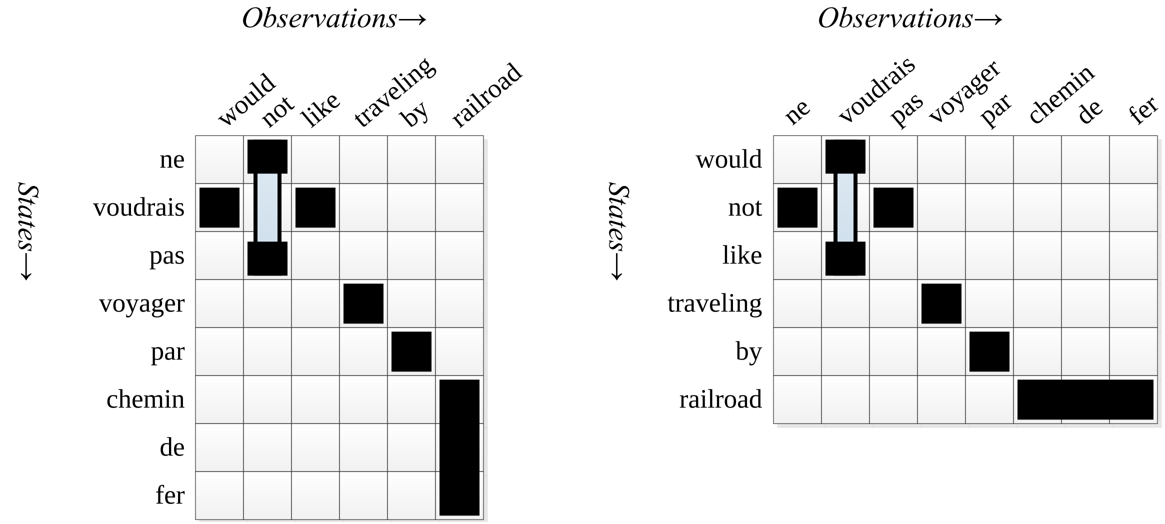

Figure 3: Example English-given-French and French-given-English alignments of the same sentence pair using the Hidden Semi-Markov Model (HSMM) for gapped-phrase-to-phrase alignment. It allows the state side phrases (denoted by vertical blocks), observation side phrases (denoted by horizontal blocks), and state-side gaps (denoted by discontinuous blocks in the same column connected by a hollow vertical "bridge"). Note both directions can capture the desired alignment for this sentence pair.

servations from the same state; in the other direction, the word *not* is generated by a single gappy phrase *ne* $\star$ *pas*. Computing posteriors for agreement is somewhat complicated, so we resort to an approximation described later. Exact inference retains a low-order polynomial runtime; we use pruning to increase speed.

2.1 Hidden Markov Alignment Models

Our model can be seen as an extension of the standard word-based Hidden Markov Model (HMM) used in alignment (Vogel et al., 1996). To ground the discussion, we first review the structure of that model. This generative model has the form $p(O|S) = \sum_A p(A, O|S)$, where $S = (s_1, \ldots, s_I) \in \Sigma^\star$ is a sequence of words from a vocabulary Σ; $O = (o_1, \ldots, o_J) \in \Pi^\star$ is a sequence from vocabulary Π; and $A = (a_1, \ldots, a_J)$ is the alignment between the two sequences. Since some words are systematically inserted during translation, the target (state) word sequence is augmented with a special NULL word. To retain the position of the last aligned word, the state space contains I copies of the NULL word, one for each position (Och and Ney, 2003). The alignment uses positive positions for words and negative positions for NULL states, so $a_j \in \{1..I\} \cup \{-1.. - I\}$, and $s_i = \text{NULL}$ if $i < 0$.

It uses the following generative procedure. First the length of the observation sequence is selected based on $p_l(J|I)$. Then for each observation position, the state is selected based on the prior state: a null state with probability p_0, or a non-null state at position a_j with probability $(1 - p_0) \cdot p_j(a_j|a_{j-1})$ where p_j is a jump distribution. Finally the observation word o_j at that position is generated with probability $p_t(o_j|s_{a_j})$, where p_t is an emission distribution:

$$p(A, O|S) = p_l(J|I) \prod_{j=1}^{J} p_j(a_j|a_{j-1}) p_t(o_j|s_{a_j})$$

$$p_j(a|a') = \begin{cases} (1 - p_0) \cdot p_d(a - |a'|) & a > 0 \\ p_0 \cdot \delta(|a|, |a'|) & a < 0 \end{cases}$$

We pick p_0 using grid search on the development set, p_l is uniform, and the p_j and p_t are optimized by EM.[2]

2.2 Gappy Semi-Markov Models

The HMM alignment model identifies a word-to-word correspondence between the observation

[2] Note that jump distances beyond -10 or 10 share a single parameter to prevent sparsity.

words and the state words. We make two changes to expand this model. First, we allow contiguous phrases on the observation side, which makes the model semi-Markov: at each time stamp, the model may emit more than one observation word. Next, we also allow contiguous and gappy phrases on the state side, leading to an alignment model that can retain phrasal links after agreement (see Section 4).

The S and O random variables are unchanged. Since a single state may generate multiple observation words, we add a new variable K representing the number of states. K should be less than J, the number of observations. The alignment variable is augmented to allow contiguous and non-contiguous ranges of words. We allow only a single gap, but of unlimited length. The null state is still present, and is again represented by negative numbers.

$$\begin{aligned} A =& (a_1, \ldots, a_K) \in \mathcal{A}(I) \\ \mathcal{A}(I) =& \{(i_1, i_2, g) | 0 < i_1 \leq i_2 \leq I, \\ & g \in \{\text{GAP}, \text{CONTIG}\}\} \cup \\ & \{(-i, -i, \text{CONTIG}) \mid 0 < i \leq I\} \end{aligned}$$

We add one more random variable to capture the total number of observations generated by each state.

$$L \in \{(l_0, l_1, \ldots, l_K) \mid 0 = l_0 < \cdots < l_K = J\}$$

The generative model takes the following form:

$$\begin{aligned} p(A, L, O|S) =& p_l(J|I) p_f(K|J) \prod_{k=1}^{K} p_j(a_k|a_{k-1}) \cdot \\ & p_t(l_k, o_{l_{k-1}+1}^{l_k} | S[a_k], l_{k-1}) \end{aligned}$$

First, the length of the observation sequence (J) is selected, based on the number of words in the state-side sentence (I). Since it does not affect the alignment, p_l is modeled as a uniform distribution. Next, we pick the total number of states to use (K), which must be less than the number of observations (J). Short state sequences receive an exponential penalty: $p_f(K|J) \propto \eta^{(J-K)}$ if $0 \leq K \leq J$, or 0 otherwise. A harsh penalty (small positive value of η) may prevent the systematic overuse of phrases.[3]

Next we decide the assignment of each state. We retain the first-order Markov assumption: the selection of each state is conditioned only on the prior state. The transition distribution is identical to the word-based HMM for single word states. For phrasal and gappy states, we jump into the first word of that state, and out of the last word of that state, and then pay a cost according to how many words are covered within that state. If $a = (i_1, i_2, g)$, then the beginning word of a is $F(a) = i_1$, the ending word is $L(a) = i_2$, and the length $N(a)$ is 2 for gapped states, 0 for null states, and $last(a) - first(a) + 1$ for all others. The transition probability is:

$$p_j(a|a') = \begin{cases} p_0 \cdot \delta(|F(a)|, |L(a')|) & \text{if } F(a) < 0 \\ (1 - p_0) p_d(F(a) - |L(a')|) \cdot & \\ \quad p_n(N(a)) & \text{otherwise} \end{cases}$$

where $p_n(c) \propto \kappa^c$ is an exponential distribution. As in the word HMM case, we use a mixture parameter p_0 to determine the likelihood of landing in a NULL state. The position of that NULL state remembers the last position of the prior state. For non-null words, we pick the first word of the state according to the distance from the last word of the prior state. Finally, we pick a length for that final state according to an exponential distribution: values of κ less than one will penalize the use of phrasal states.

For each set of state words, we maintain an emission distribution over observation word sequences. Let $S[a]$ be the set of state words referred to by the alignment variable a. For example, the English given French alignment of Figure 3 includes the following state word sets:

$$\begin{aligned} S[(2, 2, \text{CONTIG})] &= \textit{voudrais} \\ S[(1, 3, \text{GAP})] &= \textit{ne} \star \textit{pas} \\ S[(6, 8, \text{CONTIG})] &= \textit{chemin de fer} \end{aligned}$$

For the emission distribution we keep a multinomial over observation phrases for each set of state words:

$$p(l, o_{l'}^{l} | S[a], l') \propto c(o_{l'}^{l} | S[a])$$

In contrast to the approach of Deng and Byrne (2005), this encourages greater consistency across instances, and more closely resembles the commonly used phrasal translation models.

[3] We found that this penalty was crucial to prevent overfitting in independent training. Joint training with agreement made it basically unnecessary.

We note in passing that $p_f(K|J)$ may be moved inside the product: $p_f(K|J) \propto \eta^{(J-K)} = \prod_{k=1}^{K} \eta^{(l_k - l_{k-1} - 1)}$. The following form derived using the above rearrangement is helpful during EM.

$$
\begin{aligned}
p(A, L, O|S) \propto \prod_{k=1}^{K} & p_j(a_k|a_{k-1}) \cdot \\
& p_t(l_k, o_{l_{k-1}+1}^{l_k} | S[a_k], l_{k-1}) \cdot \\
& \eta^{(l_k - l_{k-1} - 1)}
\end{aligned}
$$

where $l_k - l_{k-1} - 1$ is the length of the observation phrase emitted by state $S[a_k]$.

2.3 Minimality

At alignment time we focus on finding the minimal phrase pairs, under the assumption that composed phrase pairs can be extracted in terms of these minimal pairs. We are rather strict about this, allowing only $1 \to k$ and $k \to 1$ phrasal alignment edges (or links). This should not cause undue stress, since edges of the form $2-3$ (say $e_1 e_2 \sim f_1 f_2 f_3$) can generally be decomposed into $1-1 \cup 1-2$ (i.e., $e_1 \sim f_1 \cup e_2 \sim f_2 f_3$), etc. However, the model does not require this to be true: we will describe reestimation for unconstrained general models, but use the limited form for word alignment.

3 Parameter Estimation

We use Expectation-Maximization (EM) to estimate parameters. The forward-backward algorithm efficiently computes posteriors of transitions and emissions in the word-based HMM. In a standard HMM, emission always advances the observation position by one, and the next transition is unaffected by the emission. Neither of these assumptions hold in our model: multiple observations may be emitted at a time, and a state may cover multiple state-side words, which affects the outgoing transition. A modified dynamic program computes posteriors for this generalized model.

The following formulation of the forward-backward algorithm for word-to-word alignment is a good starting point. $\alpha[x, 0, y]$ indicates the total mass of paths that have just transitioned into state y at observation x but have not yet emitted; $\alpha[x, 1, y]$ represents the mass after emission but before subsequent transition. β is defined similarly. (We omit NULL states for brevity; the extension is straightforward.)

$$
\begin{aligned}
\alpha[0, 0, y] &= p_j(y|\textsc{init}) \\
\alpha[x, 1, y] &= \alpha[x, 0, y] \cdot p_t(o_x|s_y) \\
\alpha[x, 0, y] &= \sum_{y'} \alpha[x-1, 1, y'] \cdot p_j(y|y') \\
\beta[n, 1, y] &= 1 \\
\beta[x, 0, y] &= p_t(o_x|s_y) \cdot \beta[x, 1, y] \\
\beta[x, 1, y] &= \sum_{y'} p_j(y'|y) \cdot \beta[x+1, 0, y']
\end{aligned}
$$

Not only is it easy to compute posteriors of both emissions ($\alpha[x, 0, y] p_t(o_x|s_y) \beta[x, 1, y]$) and transitions ($\alpha[x, 1, y] p_j(y'|y) \beta[x+1, 0, y']$) with this formulation, it also simplifies the generalization to complex emissions. We update the emission forward probabilities to include a search over the possible starting points in the state and observation space:

$$
\begin{aligned}
\alpha[0, 0, y] =& p_j(y|\textsc{init}) \\
\alpha[x, 1, y] =& \sum_{x' < x, y' \le y} \alpha[x', 0, y'] \cdot \textsc{emit}(x' : x, y' : y) \\
\alpha[x, 0, y] =& \sum_{y'} \alpha[x-1, 1, y'] \cdot p_j(y|y')
\end{aligned}
$$

$$
\begin{aligned}
\beta[n, 1, y] =& 1 \\
\beta[x', 0, y'] =& \sum_{x' < x, y' \le y} \textsc{emit}(x' : x, y' : y) \cdot \beta[x, 1, y] \\
\beta[x, 1, y] =& \sum_{y'} p_j(y'|y) \cdot \beta[x+1, 0, y']
\end{aligned}
$$

Phrasal and gapped emissions are pooled into EMIT:

$$
\begin{aligned}
\textsc{emit}(w : x, y : z) =& p_t(o_w^x | s_y^z) \cdot \eta^{z-y+1} \cdot \kappa^{x-w+1} + \\
& p_t(o_w^x | s_y \star s_z) \cdot \eta^2 \cdot \kappa^{x-w+1}
\end{aligned}
$$

The transition posterior is the same as above. The emission is very similar: the posterior probability that o_w^x is aligned to s_y^z is proportional to $\alpha[w, 0, y] \cdot p_t(o_w^x | s_y^z) \cdot \eta^{z-y+1} \cdot \kappa^{x-w+1} \cdot \beta[x, 1, z]$. For a gapped phrase, the posterior is proportional to $\alpha[w, 0, y] \cdot p_t(o_w^x | s_y \star s_z) \cdot \eta^2 \cdot \kappa^{x-w+1} \cdot \beta[x, 1, z]$.

Given an inference procedure for computing posteriors, unsupervised training with EM follows immediately. We use a simple maximum-likelihood update of the parameters using expected counts based on the posterior distribution.

4 Alignment by Agreement

Following Liang et al. (2006), we quantify agreement between two models as the probability that the alignments produced by the two models agree on the alignment $\mathbf{z}$ of a sentence pair $\mathbf{x} = (S, O)$:

$$\sum_{\mathbf{z}} p_1(\mathbf{z}|\mathbf{x}; \theta_1) p_2(\mathbf{z}|\mathbf{x}; \theta_2)$$

To couple the two models, the (log) probability of agreement is added to the standard log-likelihood objective:

$$\max_{\theta_1, \theta_2} \sum_{\mathbf{x}} \Big[\log p_1(\mathbf{x}; \theta_1) + \log p_2(\mathbf{x}; \theta_2) + \log \sum_{\mathbf{z}} p_1(\mathbf{z}|\mathbf{x}; \theta_1) p_2(\mathbf{z}|\mathbf{x}; \theta_2) \Big]$$

We use the heuristic estimator from Liang et al. (2006), letting q be a product of marginals:

$$\mathrm{E} : q(\mathbf{z}; \mathbf{x}) := \prod_{z \in \mathbf{z}} p_1(z|\mathbf{x}; \theta_1) p_2(z|\mathbf{x}; \theta_2)$$

where each $p_k(z|\mathbf{x}; \theta_k)$ is the posterior marginal of some edge z according to each model. Such a heuristic E step computes the marginals for each model separately, then multiplies the marginals corresponding to the same edge. This product of marginals acts as the approximation to the posterior used in the M step for each model. The intuition is that if the two models disagree on a certain edge z, then the marginal product is small, hence that edge is dis-preferred in each model.

Contiguous phrase agreement. It is simple to extend agreement to alignments in the absence of gaps. Multi-word (phrasal) links are assigned some posterior probability in both models, as shown in the example in Figure 3, and we multiply the posteriors of these phrasal links just as in the single word case.[4]

$$\begin{aligned}\gamma_{F\to E}(f_i, e_j) &:= \gamma_{E\to F}(e_j, f_i) \\ &:= [\gamma_{F\to E}(f_i, e_j) \times \gamma_{E\to F}(e_j, f_i)]\end{aligned}$$

Gappy phrase agreement. When we introduce gappy phrasal states, agreement becomes more challenging. In the forward direction F→E, if we have a gappy state aligned to an observation, say $f_i \star f_j \sim e_k$, then its corresponding edge in the backward direction E→F would be $e_k \backsim f_i \star f_j$. However, this is represented by two distinct and unrelated emissions. Although it is possible the compute the posterior probability of two non-adjacent emissions, this requires running a separate dynamic program for each such combination to sum the mass between these emissions. For the sake of efficiency we resort to an approximate computation of posterior marginals using the two word-to-word edges $e_k \backsim f_i$ and $e_k \backsim f_j$.

The forward posterior $\gamma_{F\to E}$ for edge $f_i \star f_j \sim e_k$ is multiplied with the $\min$ of the backward posteriors of the edges $e_k \backsim f_i$ and $e_k \backsim f_j$.

$$\begin{aligned}\gamma_{F\to E}(f_i \star f_j, e_k) &:= \gamma_{F\to E}(f_i \star f_j, e_k) \times \\ &\min \Big\{ \gamma_{E\to F}(e_k, f_i), \gamma_{E\to F}(e_k, f_j) \Big\}\end{aligned}$$

Note that this $\min$ is an upper bound on the desired posterior of edge $e_k \backsim f_i \star f_j$, since every path that passes through $e_k \backsim f_i$ and $e_k \backsim f_j$ must pass through $e_k \backsim f_i$, therefore the posterior of $e_k \backsim f_i \star f_j$ is less than that of $e_k \backsim f_i$, and likewise less than that of $e_k \backsim f_j$.

The backward posteriors of the edges $e_k \backsim f_i$ and $e_k \backsim f_j$ are also mixed with the forward posteriors of the edges to which they correspond.

$$\gamma_{E\to F}(e_k, f_i) := \gamma_{E\to F}(e_k, f_i) \times \Bigg[\gamma_{F\to E}(f_i, e_k) + \sum_{h<i<j} \Big\{ \gamma_{F\to E}(f_h \star f_i, e_k) + \gamma_{F\to E}(f_i \star f_j, e_k) \Big\} \Bigg]$$

5 Pruned Lists of 'Allowed' Phrases

To identify contiguous and gapped phrases that are more likely to lead to good alignments, we use word-to-word HMM alignments from the full training data in both directions (F→E and E→F). We collect observation phrases of length 2 to K aligned to a single state, i.e. $o_i^j \sim s$, to add to a list of allowed phrases. For gappy phrases, we find all non-consecutive observation pairs o_i and o_j such that: (a) both are

[4] Phrasal correspondences can be represented in multiple ways: multiple adjacent words could be generated from the same state either using one semi-Markov emission, or using multiple single word emissions followed by self-jumps. Only the first case is reinforced through agreement, so the latter is implicitly discouraged. We explored an option to forbid same-state transitions, but found it made little difference in practice.

aligned to the same state s_k, (b) state s_k is aligned to only these two observations, and (c) at least one observation between o_i and o_j is aligned to a non-null state other than s_k. These observation phrases are collected from F→E and E→F models to build contiguous and gappy phrase lists for both languages.

Next, we order the phrases in each contiguous list using the discounted probability:

$$p_\delta(o_i^j \sim s|o_i^j) = \frac{\max(0, count(o_i^j \sim s) - \delta)}{count(o_i^j)}$$

where $count(o_i^j \sim s)$ is the count of occurrence of the observation-phrase o_i^j, all aligned to some single state s, and $count(o_i^j)$ is the count of occurrence of the observation phrase o_i^j, not all necessarily aligned to a single state. Similarly, we rank the gappy phrases using the discounted probability:

$$p_\delta(o_i \star o_j \sim s|o_i \star o_j) = \frac{\max(0, count(o_i \star o_j \sim s) - \delta)}{count(o_i \star o_j)}$$

where $count(o_i \star o_j \sim s)$ is the count of occurrence of the observations o_i and o_j aligned to a single state s with the conditions mentioned above, and $count(o_i \star o_j)$ is the count of general occurrence of the observations o_i and o_j in order. We find that 200 gappy phrases and 1000 contiguous phrases works well, based on tuning with a development set.

6 Complexity Analysis

Let m be the length of the state sentence S and n be the length of the observation sentence O. In IBM Model 1 (Brown et al., 1994), with only a translation model, we can infer posteriors or max alignments in $\mathcal{O}(mn)$. HMM-based word-to-word alignment model (Vogel et al., 1996) adds a distortion model, increasing the complexity to $\mathcal{O}(m^2n)$.

Introducing phrases (contiguous) on the observation side, we get a HSMM (Hidden Semi-Markov Model). If we allow phrases of length no greater than K, then the number of observation types rises from n to Kn for an overall complexity of $\mathcal{O}(m^2Kn)$. Introducing state phrases (contiguous) with length $\leq K$ grows the number of state types from m to Km. Complexity further increases to $\mathcal{O}((Km)^2Kn) = \mathcal{O}(K^3m^2n)$.

Finally, when we introduce gappy state phrases of the type $s_i \star s_j$, the number of such phrases is $\mathcal{O}(m^2)$, since we may choose a start and end point independently. Thus, the total complexity rises to $\mathcal{O}((Km + m^2)^2Kn) = \mathcal{O}(Km^4n)$. Although this is less than the $\mathcal{O}(n^6)$ complexity of exact ITG (Inversion Transduction Grammar) model (Wu, 1997), a quintic algorithm is often quite slow.

The pruned lists of allowed phrases limit this complexity. The model is allowed to use observation (contiguous) and state (contiguous and gappy) phrases only from these lists. The number of phrases that match any given sentence pair from these pruned lists is very small ($\sim$ 2 to 5). If the number of phrases in the lists that match the observation and state side of a given sentence pair are small constants, the complexity remains $\mathcal{O}(m^2n)$, equal to that of word-based models.

7 Results

We evaluate our models based on both word alignment and end-to-end translation with two language pairs: English-French and English-German. For French-English, we use the Hansards NAACL 2003 shared-task dataset, which contains nearly 1.1 million training sentence pairs. We also evaluated on German-English Europarl data from WMT2010, with nearly 1.6 million training sentence pairs. The model from Liang et al. (2006) is our word-based baseline.

7.1 Training Regimen

Our training regimen begins with both the forward (F→E) and backward (E→F) iterations of Model 1 run independently (i.e. without agreement). Next, we train several iterations of the forward and backward word-to-word HMMs, again with independent training. We do not use agreement during word alignment since it tends to produce sparse 1-1 alignments, which in turn leads to low phrase emission probabilities in the gappy model.

Initializing the emission probabilities of the semi-Markov model is somewhat complicated, since the word-based models do not assign any mass to the phrasal or gapped configurations. Therefore we use a heuristic method. We first retrieve the Viterbi alignments of the forward and backward word-to-word HMM aligners. For phrasal correspondences, we combine these forward and backward Viterbi alignments using a common heuristic (Union, Intersection, Refined, or Grow-Diag-Final), and extract tight phrase-pairs (no unaligned words on the boundary) from this alignment set. We found that Grow-Diag-Final was most effective in our experiments. The counts gathered from this phrase extraction are used to initialize phrasal translation probabilities. For gappy states in a forward (F→E) model, we use alignments from the backward (E→F) model. If a state s_k is aligned to two non-consecutive observations o_i and o_j such that s_k is not aligned to any other observation, and at least one observation between o_i and o_j is aligned to a non-null state other than s_k, then we reverse this link to get $o_i \ \star \ o_j \sim s_k$ and use it as a gapped-state-phrase instance for adding fractional counts. Given these approximate fractional counts, we perform a standard MLE M-step to initialize the emission probability distributions. The distortion probabilities from the word-based model are used without changes.

7.2 Alignment Results (F1)

The validation and test sentences have been hand-aligned (see Och and Ney (2003)) and are marked with both sure and possible alignments. For French-English, following Liang et al. (2006), we lowercase all words, and use the validation set plus the first 100 test sentences as our development set and the remaining 347 test-sentences as our test-set for final F1 evaluation.[5] In German-English, we have a development set of 102 sentences, and a test set of 258 sentences, also annotated with a set of sure and possible alignments. Given a predicted alignment A, precision and recall are computed using sure alignments S and possible alignments P (where $S \subseteq P$) as in Och and Ney (2003):

$$Precision = \frac{|A \cap P|}{|A|} \times 100\%$$

$$Recall = \frac{|A \cap S|}{|S|} \times 100\%$$

Language pair	Word-to-word	Gappy
French-English	34.0	34.5
German-English	19.3	19.8

Table 2: BLEU results on German-English and French-English.

$$AER = \left(1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|}\right) \times 100\%$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \times 100\%$$

Many free parameters were tuned to optimize alignment F1 on the development set, including the number of iterations of each Model 1, HMM, and Gappy; the NULL weight p_0, the number of contiguous and gappy phrases to include, and the maximum phrase length. Five iterations of all models, $p_0 = 0.3$, using the top 1000 contiguous phrases and the top 200 gappy phrases, maximum phrase length of 5, and penalties $\eta = \kappa = 1$ produced competitive results. Note that by setting η and κ to one, we have effectively removed the penalty altogether without affecting our results. In Table 1 we see a consistent improvement with the addition of contiguous phrases, and some additional gains with gappy phrases.

7.3 Translation Results (BLEU)

We assembled a phrase-based system from the alignments (using only contiguous phrases consistent with the potentially gappy alignment), with 4 channel models, word and phrase count features, distortion penalty, lexicalized reordering model, and a 5-gram language model, weighted by MERT. The same free parameters from above were tuned to optimize development set BLEU using grid search. The improvements in Table 2 are encouraging, especially as a syntax-based or non-contiguous phrasal system (Galley and Manning, 2010) may benefit more from gappy phrases.

8 Conclusions and Future Work

We have described an algorithm for efficient unsupervised alignment of phrases. Relatively straightforward extensions to the base HMM allow for efficient inference, and agreement between the two

[5] We report F1 rather than AER because AER appears not to correlate well with translation quality.(Fraser and Marcu, 2007)

Data	Decoding method	Word-to-word	+Contig phrases	+Gappy phrases
FE 10K	Viterbi	89.7	90.6	90.3
FE 10K	Posterior ≥ 0.1	90.1	90.4	**90.7**
FE 100K	Viterbi	93.0	93.6	93.8
FE 100K	Posterior ≥ 0.1	93.1	93.7	**93.8**
FE All	Viterbi	94.1	94.3	94.3
FE All	Posterior ≥ 0.1	94.2	94.4	**94.5**
GE 10K	Viterbi	76.2	79.6	**79.7**
GE 10K	Posterior ≥ 0.1	76.7	79.3	79.3
GE 100K	Viterbi	81.0	83.0	83.2
GE 100K	Posterior ≥ 0.1	80.7	83.1	**83.4**
GE All	Viterbi	83.0	85.2	85.6
GE All	Posterior ≥ 0.1	83.7	85.3	**85.7**

Table 1: F1 scores of automatic word alignments, evaluated on the test set of the hand-aligned sentence pairs.

models prevents EM from overfitting, even in the absence of harsh penalties. We also allow gappy (non-contiguous) phrases on the state side, which makes agreement more successful but agreement needs approximation of posterior marginals. Using pruned lists of good phrases, we maintain complexity equal to the baseline word-to-word model.

There are several steps forward from this point. Limiting the gap length also prevents combinatorial explosion; we hope to explore this in future work. Clearly a translation system that uses discontinuous mappings at runtime (Chiang, 2007; Galley and Manning, 2010) may make better use of discontinuous alignments. This model can also be applied at the morpheme or character level, allowing joint inference of segmentation and alignment. Furthermore the state space could be expanded and enhanced to include more possibilities: states with multiple gaps might be useful for alignment in languages with template morphology, such as Arabic or Hebrew. More exploration in the model space could be useful – a better distortion model might place a stronger distribution on the likely starting and ending points of phrases.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful suggestions. This project is funded by Microsoft Research.

References

Jesús Andrés-Ferrer and Alfons Juan. 2009. A phrase-based hidden semi-Markov approach to machine translation. In *Proceedings of EAMT*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*.

Hal Daumé III and Daniel Marcu. 2004. A phrase-based HMM approach to document/abstract alignment. In *Proceedings of EMNLP*.

John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings of ACL*.

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

Xiangyu Duan, Min Zhang, and Haizhou Li. 2010. Pseudo-word for phrase-based machine translation. In *Proceedings of ACL*.

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

Michel Galley and Christopher D. Manning. 2010. Accurate non-hierarchical phrase-based translation. In *HLT/NAACL*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of HLT-NAACL*.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proceedings of HLT-NAACL*.

Patrik Lambert and Rafael Banchs. 2006. Grouping multi-word expressions according to part-of-speech in statistical machine translation. In *Proc. of the EACL Workshop on Multi-Word-Expressions in a Multilingual Context*.

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

Yanjun Ma, Nicolas Stroppa, and Andy Way. 2007. Boostrapping word alignment via word packing. In *Proceedings of ACL.*

Daniel Marcu and Daniel Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of EMNLP*.

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

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of COLING*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. 2006. Syntax augmented machine translation via chart parsing. In *Processings of the Statistical Machine Translation Workshop at NAACL.*