Discriminative Modeling of Extraction Sets for Machine Translation

John DeNero and **Dan Klein**
Computer Science Division
University of California, Berkeley
{denero,klein}@cs.berkeley.edu

Abstract

We present a discriminative model that directly predicts which set of phrasal translation rules should be extracted from a sentence pair. Our model scores *extraction sets*: nested collections of all the overlapping phrase pairs consistent with an underlying word alignment. Extraction set models provide two principle advantages over word-factored alignment models. First, we can incorporate features on phrase pairs, in addition to word links. Second, we can optimize for an extraction-based loss function that relates directly to the end task of generating translations. Our model gives improvements in alignment quality relative to state-of-the-art unsupervised and supervised baselines, as well as providing up to a 1.4 improvement in BLEU score in Chinese-to-English translation experiments.

1 Introduction

In the last decade, the field of statistical machine translation has shifted from generating sentences word by word to systems that recycle whole fragments of training examples, expressed as *translation rules*. This general paradigm was first pursued using contiguous phrases (Och et al., 1999; Koehn et al., 2003), and has since been generalized to a wide variety of hierarchical and syntactic formalisms. The training stage of statistical systems focuses primarily on discovering translation rules in parallel corpora.

Most systems discover translation rules via a two-stage pipeline: a parallel corpus is aligned at the word level, and then a second procedure extracts fragment-level rules from word-aligned sentence pairs. This paper offers a model-based alternative to phrasal rule extraction, which merges this two-stage pipeline into a single step. We present a discriminative model that directly predicts which set of phrasal translation rules should be extracted from a sentence pair. Our model predicts *extraction sets*: combinatorial objects that include the set of all overlapping phrasal translation rules consistent with an underlying word-level alignment. This approach provides additional discriminative power relative to word aligners because extraction sets are scored based on the phrasal rules they contain in addition to word-to-word alignment links. Moreover, the structure of our model directly reflects the purpose of alignment models in general, which is to discover translation rules.

We address several challenges to training and applying an extraction set model. First, we would like to leverage existing word-level alignment resources. To do so, we define a deterministic mapping from word alignments to extraction sets, inspired by existing extraction procedures. In our mapping, *possible* alignment links have a precise interpretation that dictates what phrasal translation rules can be extracted from a sentence pair. This mapping allows us to train with existing annotated data sets and use the predictions from word-level aligners as features in our extraction set model.

Second, our model solves a structured prediction problem, and the choice of loss function during training affects model performance. We optimize for a phrase-level F-measure in order to focus learning on the task of predicting phrasal rules rather than word alignment links.

Third, our discriminative approach requires that we perform inference in the space of extraction sets. Our model does not factor over disjoint word-to-word links or minimal phrase pairs, and so existing inference procedures do not directly apply. However, we show that the dynamic program for a block ITG aligner can be augmented to score extraction sets that are indexed by underlying ITG word alignments (Wu, 1997). We also describe a

Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1453–1463,
Uppsala, Sweden, 11-16 July 2010.

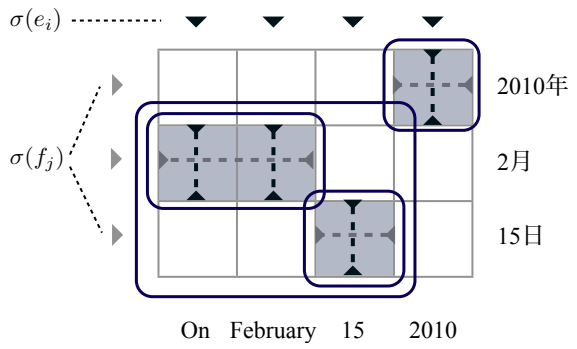

Figure 1: A word alignment $\mathcal{A}$ (shaded grid cells) defines projections $\sigma(e_i)$ and $\sigma(f_j)$, shown as dotted lines for each word in each sentence. The extraction set $R_3(\mathcal{A})$ includes all bispans licensed by these projections, shown as rounded rectangles.

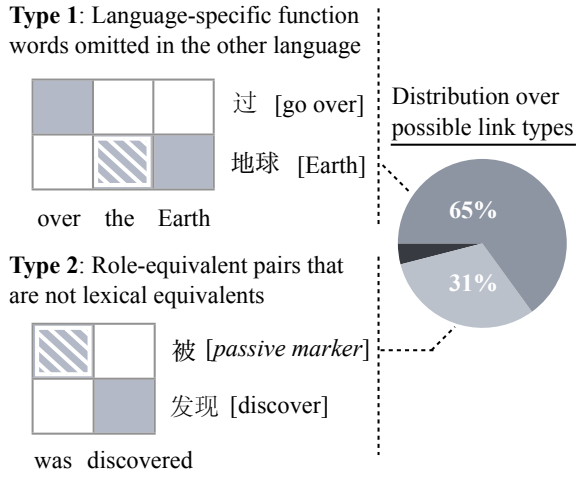

Figure 2: Examples of two types of possible alignment links (striped). These types account for 96% of the possible alignment links in our data set.

coarse-to-fine inference approach that allows us to scale our method to long sentences.

Our extraction set model outperforms both unsupervised and supervised word aligners at predicting word alignments and extraction sets. We also demonstrate that extraction sets are useful for end-to-end machine translation. Our model improves translation quality relative to state-of-the-art Chinese-to-English baselines across two publicly available systems, providing total BLEU improvements of 1.2 in Moses, a phrase-based system, and 1.4 in a Joshua, a hierarchical system (Koehn et al., 2007; Li et al., 2009)

2 Extraction Set Models

The input to our model is an unaligned sentence pair, and the output is an extraction set of phrasal translation rules. Word-level alignments are generated as a byproduct of inference. We first specify the relationship between word alignments and extraction sets, then define our model.

2.1 Extraction Sets from Word Alignments

Rule extraction is a standard concept in machine translation: word alignment constellations license particular sets of overlapping rules, from which subsets are selected according to limits on phrase length (Koehn et al., 2003), number of gaps (Chiang, 2007), count of internal tree nodes (Galley et al., 2006), etc. In this paper, we focus on phrasal rule extraction (i.e., phrase pair extraction), upon which most other extraction procedures are based.

Given a sentence pair $(\mathbf{e}, \mathbf{f})$, phrasal rule extraction defines a mapping from a set of word-to-word alignment links $\mathcal{A} = \{(i, j)\}$ to an *extraction set* of bispans $R_n(\mathcal{A}) = \{[g, h) \Leftrightarrow [k, \ell)\}$, where each bispan links target span $[g, h)$ to source span $[k, \ell)$.[1] The maximum phrase length n ensures that $\max(h - g, \ell - k) \leq n$.

We can describe this mapping via word-to-phrase projections, as illustrated in Figure 1. Let word e_i project to the phrasal span $\sigma(e_i)$, where

$$\sigma(e_i) = \left[\min_{j \in J_i} j \ , \ \max_{j \in J_i} j + 1 \right) \qquad (1)$$
$$J_i = \{j : (i, j) \in \mathcal{A}\}$$

and likewise each word f_j projects to a span of $\mathbf{e}$. Then, $R_n(\mathcal{A})$ includes a bispan $[g, h) \Leftrightarrow [k, \ell)$ iff

$$\sigma(e_i) \subseteq [k, \ell) \qquad \forall i \in [g, h)$$
$$\sigma(f_j) \subseteq [g, h) \qquad \forall j \in [k, \ell)$$

That is, every word in one of the phrasal spans must project within the other. This mapping is deterministic, and so we can interpret a word-level alignment $\mathcal{A}$ as also specifying the phrasal rules that should be extracted from a sentence pair.

2.2 Possible and Null Alignment Links

We have not yet accounted for two special cases in annotated corpora: *possible* alignments and *null* alignments. To analyze these annotations, we consider a particular data set: a hand-aligned portion

[1] We use the fencepost indexing scheme used commonly for parsing. Words are 0-indexed. Spans are inclusive on the lower bound and exclusive on the upper bound. For example, the span $[0, 2)$ includes the first two words of a sentence.

of the NIST MT02 Chinese-to-English test set, which has been used in previous alignment experiments (Ayan et al., 2005; DeNero and Klein, 2007; Haghighi et al., 2009).

Possible links account for 22% of all alignment links in these data, and we found that most of these links fall into two categories. First, possible links are used to align function words that have no equivalent in the other language, but colocate with aligned content words, such as English determiners. Second, they are used to mark pairs of words or short phrases that are not lexical equivalents, but which play equivalent roles in each sentence. Figure 2 shows examples of these two use cases, along with their corpus frequencies.[2]

On the other hand, null alignments are used sparingly in our annotated data. More than 90% of words participate in some alignment link. The unaligned words typically express content in one sentence that is absent in its translation.

Figure 3 illustrates how we interpret possible and null links in our projection. Possible links are typically not included in extraction procedures because most aligners predict only sure links. However, we see a natural interpretation for possible links in rule extraction: they license phrasal rules that both include and exclude them. We exclude null alignments from extracted phrases because they often indicate a mismatch in content.

We achieve these effects by redefining the projection operator σ. Let $\mathcal{A}^{(s)}$ be the subset of $\mathcal{A}$ that are *sure* links, then let the index set J_i used for projection σ in Equation 1 be

$$J_i = \begin{cases} \{j : (i,j) \in \mathcal{A}^{(s)}\} & \text{if } \exists j : (i,j) \in \mathcal{A}^{(s)} \\ \{-1, |\mathbf{f}|\} & \text{if } \nexists j : (i,j) \in \mathcal{A} \\ \{j : (i,j) \in \mathcal{A}\} & \text{otherwise} \end{cases}$$

Here, J_i is a set of integers, and $\sigma(e_i)$ for null aligned e_i will be $[-1, |\mathbf{f}| + 1)$ by Equation 1.

Of course, the characteristics of our aligned corpus may not hold for other annotated corpora or other language pairs. However, we hope that the overall effectiveness of our modeling approach will influence future annotation efforts to build corpora that are consistent with this interpretation.

2.3 A Linear Model of Extraction Sets

We now define a linear model that scores extraction sets. We restrict our model to score only *coherent* extraction sets $R_n(\mathcal{A})$, those that are licensed by an underlying word alignment $\mathcal{A}$ with sure alignments $\mathcal{A}^{(s)} \subseteq \mathcal{A}$. Conditioned on a sentence pair $(\mathbf{e}, \mathbf{f})$ and maximum phrase length n, we score extraction sets via a feature vector $\phi(\mathcal{A}^{(s)}, R_n(\mathcal{A}))$ that includes features on sure links $(i,j) \in \mathcal{A}^{(s)}$ and features on the bispans in $R_n(\mathcal{A})$ that link $[g,h)$ in $\mathbf{e}$ to $[k,\ell)$ in $\mathbf{f}$:

$$\phi(\mathcal{A}^{(s)}, R_n(\mathcal{A})) = \sum_{(i,j)\in\mathcal{A}^{(s)}} \phi_a(i,j) + \sum_{[g,h)\Leftrightarrow[k,\ell)\in R_n(\mathcal{A})} \phi_b(g,h,k,\ell)$$

Because the projection operator $R_n(\cdot)$ is a deterministic function, we can abbreviate $\phi(\mathcal{A}^{(s)}, R_n(\mathcal{A}))$ as $\phi(\mathcal{A})$ without loss of information, although we emphasize that $\mathcal{A}$ is a set of sure and possible alignments, and $\phi(\mathcal{A})$ does not decompose as a sum of vectors on individual word-level alignment links. Our model is parameterized by a weight vector θ, which scores an extraction set $R_n(\mathcal{A})$ as $\theta \cdot \phi(\mathcal{A})$.

To further limit the space of extraction sets we are willing to consider, we restrict $\mathcal{A}$ to block inverse transduction grammar (ITG) alignments, a space that allows many-to-many alignments through phrasal terminal productions, but otherwise enforces at-most-one-to-one phrase matchings with ITG reordering patterns (Cherry and Lin, 2007; Zhang et al., 2008). The ITG constraint

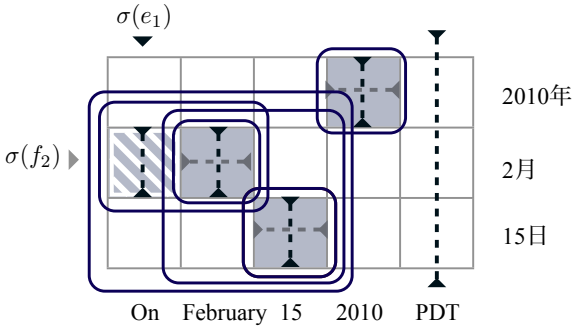

Figure 3: Possible links constrain the word-to-phrase projection of otherwise unaligned words, which in turn license overlapping phrases. In this example, $\sigma(f_2) = [1,2)$ does not include the possible link at $(1,0)$ because of the sure link at $(1,1)$, but $\sigma(e_1) = [1,2)$ does use the possible link because it would otherwise be unaligned. The word "PDT" is null aligned, and so its projection $\sigma(e_4) = [-1,4)$ extends beyond the bounds of the sentence, excluding "PDT" from all phrase pairs.

[2]We collected corpus frequencies of possible alignment link types ourselves on a sample of the hand-aligned data set.

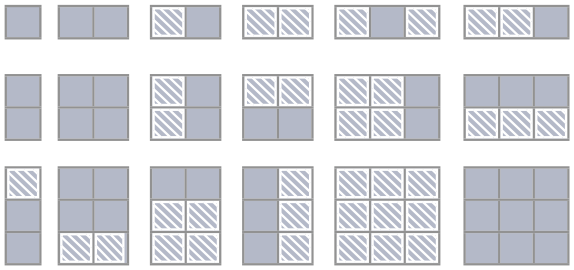

Figure 4: Above, we show a representative subset of the block alignment patterns that serve as terminal productions of the ITG that restricts the output space of our model. These terminal productions cover up to $n = 3$ words in each sentence and include a mixture of sure (filled) and possible (striped) word-level alignment links.

is more computationally convenient than arbitrarily ordered phrase matchings (Wu, 1997; DeNero and Klein, 2008). However, the space of block ITG alignments is expressive enough to include the vast majority of patterns observed in hand-annotated parallel corpora (Haghighi et al., 2009).

In summary, our model scores all $R_n(\mathcal{A})$ for $\mathcal{A} \in \text{ITG}(\mathbf{e}, \mathbf{f})$ where $\mathcal{A}$ can include block terminals of size up to n. In our experiments, $n = 3$. Unlike previous work, we allow possible alignment links to appear in the block terminals, as depicted in Figure 4.

3 Model Estimation

We estimate the weights θ of our extraction set model discriminatively using the margin-infused relaxed algorithm (MIRA) of Crammer and Singer (2003)—a large-margin, perceptron-style, online learning algorithm. MIRA has been used successfully in MT to estimate both alignment models (Haghighi et al., 2009) and translation models (Chiang et al., 2008).

For each training example, MIRA requires that we find the alignment $\mathcal{A}_m$ corresponding to the highest scoring extraction set $R_n(\mathcal{A}_m)$ under the current model,

$$\mathcal{A}_m = \arg\max_{\mathcal{A} \in \text{ITG}(\mathbf{e},\mathbf{f})} \theta \cdot \phi(\mathcal{A}) \qquad (2)$$

Section 4 describes our approach to solving this search problem for model inference.

MIRA updates away from $R_n(A_m)$ and toward a gold extraction set $R_n(A_g)$. Some hand-annotated alignments are outside of the block ITG model class. Hence, we update toward the extraction set for a pseudo-gold alignment $\mathcal{A}_g \in \text{ITG}(\mathbf{e}, \mathbf{f})$ with minimal distance from the true reference alignment $\mathcal{A}_t$.

$$\mathcal{A}_g = \arg\min_{\mathcal{A} \in \text{ITG}(\mathbf{e},\mathbf{f})} |\mathcal{A} \cup \mathcal{A}_t - \mathcal{A} \cap \mathcal{A}_t| \quad (3)$$

Inference details appear in Section 4.3.

Given $\mathcal{A}_g$ and $\mathcal{A}_m$, we update the model parameters away from $\mathcal{A}_m$ and toward $\mathcal{A}_g$.

$$\theta \leftarrow \theta + \tau \cdot (\phi(\mathcal{A}_g) - \phi(\mathcal{A}_m))$$

where τ is the minimal step size that will ensure we prefer $\mathcal{A}_g$ to $\mathcal{A}_m$ by a margin greater than the loss $L(\mathcal{A}_m; \mathcal{A}_g)$, capped at some maximum update size C to provide regularization. We use $C = 0.01$ in experiments. The step size is a closed form function of the loss and feature vectors: $\tau =$

$$\min\left(C, \frac{L(\mathcal{A}_m; \mathcal{A}_g) - \theta \cdot (\phi(\mathcal{A}_g) - \phi(\mathcal{A}_m))}{||\phi(\mathcal{A}_g) - \phi(\mathcal{A}_m)||_2^2}\right)$$

We train the model for 30 iterations over the training set, shuffling the order each time, and we average the weight vectors observed after each iteration to estimate our final model.

3.1 Extraction Set Loss Function

In order to focus learning on predicting the right bispans, we use an extraction-level loss $L(\mathcal{A}_m; \mathcal{A}_g)$: an F-measure of the overlap between bispans in $R_n(\mathcal{A}_m)$ and $R_n(\mathcal{A}_g)$. This measure has been proposed previously to evaluate alignment systems (Ayan and Dorr, 2006). Based on preliminary translation results during development, we chose bispan F_5 as our loss:

$$\text{Pr}(\mathcal{A}_m) = |R_n(\mathcal{A}_m) \cap R_n(\mathcal{A}_g)|/|R_n(\mathcal{A}_m)|$$
$$\text{Rc}(\mathcal{A}_m) = |R_n(\mathcal{A}_m) \cap R_n(\mathcal{A}_g)|/|R_n(\mathcal{A}_g)|$$
$$\text{F}_5(\mathcal{A}_m; \mathcal{A}_g) = \frac{(1+5^2) \cdot \text{Pr}(\mathcal{A}_m) \cdot \text{Rc}(\mathcal{A}_m)}{5^2 \cdot \text{Pr}(\mathcal{A}_m) + \text{Rc}(\mathcal{A}_m)}$$
$$L(\mathcal{A}_m; \mathcal{A}_g) = 1 - \text{F}_5(\mathcal{A}_m; \mathcal{A}_g)$$

F_5 favors recall over precision. Previous alignment work has shown improvements from adjusting the F-measure parameter (Fraser and Marcu, 2006). In particular, Lacoste-Julien et al. (2006) also chose a recall-biased objective.

Optimizing for a bispan F-measure penalizes alignment mistakes in proportion to their rule extraction consequences. That is, adding a word link that prevents the extraction of many correct phrasal rules, or which licenses many incorrect rules, is strongly discouraged by this loss.

3.2 Features on Extraction Sets

The discriminative power of our model is driven by the features on sure word alignment links $\phi_a(i, j)$ and bispans $\phi_b(g, h, k, \ell)$. In both cases, the most important features come from the predictions of unsupervised models trained on large parallel corpora, which provide frequency and co-occurrence information.

To score word-to-word links, we use the posterior predictions of a jointly trained HMM alignment model (Liang et al., 2006). The remaining features include a dictionary feature, an identical word feature, an absolute position distortion feature, and features for numbers and punctuation.

To score phrasal translation rules in an extraction set, we use a mixture of feature types. Extraction set models allow us to incorporate the same phrasal relative frequency statistics that drive phrase-based translation performance (Koehn et al., 2003). To implement these frequency features, we extract a phrase table from the alignment predictions of a jointly trained unsupervised HMM model using Moses (Koehn et al., 2007), and score bispans using the resulting features. We also include indicator features on lexical templates for the 50 most common words in each language, as in Haghighi et al. (2009). We include indicators for the number of words and Chinese characters in rules. One useful indicator feature exploits the fact that capitalized terms in English tend to align to Chinese words with three or more characters. On 1-by-n or n-by-1 phrasal rules, we include indicator features of fertility for common words.[3]

We also include monolingual phrase features that expose useful information to the model. For instance, English bigrams beginning with "the" are often extractable phrases. English trigrams with a hyphen as the second word are typically extractable, meaning that the first and third words align to consecutive Chinese words. When any conjugation of the word "to be" is followed by a verb, indicating passive voice or progressive tense, the two words tend to align together.

Our feature set also includes bias features on phrasal rules and links, which control the number of null-aligned words and number of rules licensed. In total, our final model includes 4,249 individual features, dominated by various instantiations of lexical templates.

[3]Limiting lexicalized features to common words helps prevent overfitting.

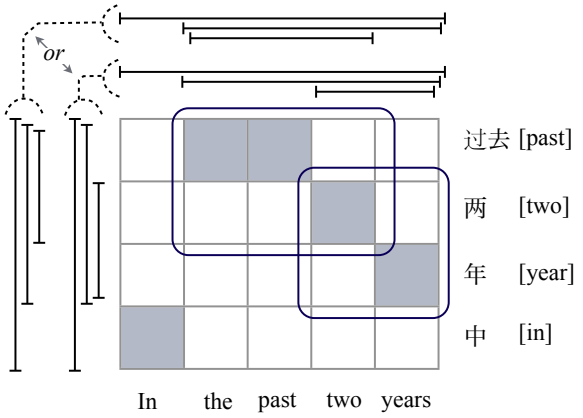

Figure 5: Both possible ITG decompositions of this example alignment will split one of the two highlighted bispans across constituents.

4 Model Inference

Equation 2 asks for the highest scoring extraction set under our model, $R_n(\mathcal{A}_m)$, which we also require at test time. Although we have restricted $\mathcal{A}_m \in \text{ITG}(\mathbf{e}, \mathbf{f})$, our extraction set model does not factor over ITG productions, and so the dynamic program for a vanilla block ITG will not suffice to find $R_n(\mathcal{A}_m)$. To see this, consider the extraction set in Figure 5. An ITG decomposition of the underlying alignment imposes a hierarchical bracketing on each sentence, and some bispan in the extraction set for this alignment will cross any such bracketing. Hence, the score of some licensed bispan will be non-local to the ITG decomposition.

4.1 A Dynamic Program for Extraction Sets

If we treat the maximum phrase length n as a fixed constant, then we can define a dynamic program to search the space of extraction sets. An ITG derivation for some alignment $\mathcal{A}$ decomposes into two sub-derivations for $\mathcal{A}_L$ and $\mathcal{A}_R$.[4] The model score of $\mathcal{A}$, which scores extraction set $R_n(\mathcal{A})$, decomposes over $\mathcal{A}_L$ and $\mathcal{A}_R$, along with any phrasal bispans licensed by adjoining $\mathcal{A}_L$ and $\mathcal{A}_R$.

$$\theta \cdot \phi(\mathcal{A}) = \theta \cdot \phi(\mathcal{A}_L) + \theta \cdot \phi(\mathcal{A}_R) + I(\mathcal{A}_L, \mathcal{A}_R)$$

where $I(\mathcal{A}_L, \mathcal{A}_R)$ is $\theta \cdot \sum \phi(g, h, k, l)$ summed over licensed bispans $[g, h) \Leftrightarrow [k, \ell)$ that overlap the boundary between $\mathcal{A}_L$ and $\mathcal{A}_R$.[5]

[4]We abuse notation in conflating an alignment $\mathcal{A}$ with its derivation. All derivations of the same alignment receive the same score, and we only compute the max, not the sum.

[5]We focus on the case of adjoining two aligned bispans. Our algorithm easily extends to include null alignments, but we focus on the non-null setting for simplicity.

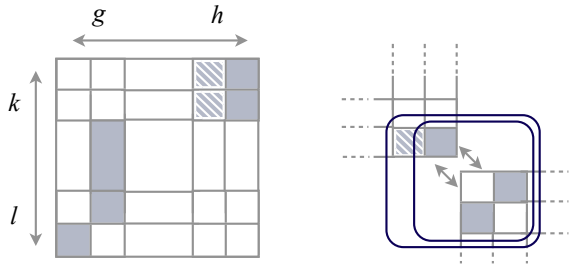

Figure 6: Augmenting the ITG grammar states with the alignment configuration in an $n-1$ deep perimeter of the bispan allows us to score all overlapping phrasal rules introduced by adjoining two bispans. The state must encode whether a sure link appears in each edge column or row, but the specific location of edge links is not required.

In order to compute $I(\mathcal{A}_L, \mathcal{A}_R)$, we need certain information about the alignment configurations of $\mathcal{A}_L$ and $\mathcal{A}_R$ where they adjoin at a corner. The state must represent (a) the specific alignment links in the $n-1$ deep corner of each $\mathcal{A}$, and (b) whether any sure alignments appear in the rows or columns extending from those corners.[6] With this information, we can infer the bispans licensed by adjoining $\mathcal{A}_L$ and $\mathcal{A}_R$, as in Figure 6.

Applying our score recurrence yields a polynomial-time dynamic program. This dynamic program is an instance of ITG bitext parsing, where the grammar uses symbols to encode the alignment contexts described above. This context-as-symbol augmentation of the grammar is similar in character to augmenting symbols with lexical items to score language models during hierarchical decoding (Chiang, 2007).

4.2 Coarse-to-Fine Inference and Pruning

Exhaustive inference under an ITG requires $O(k^6)$ time in sentence length k, and is prohibitively slow when there is no sparsity in the grammar. Maintaining the context necessary to score non-local bispans further increases running time. That is, ITG inference is organized around search states associated with a grammar symbol and a bispan; augmenting grammar symbols also augments this state space.

To parse quickly, we prune away search states using predictions from the more efficient HMM alignment model (Ney and Vogel, 1996). We discard all states corresponding to bispans that are incompatible with 3 or more alignment links under an intersected HMM—a proven approach to pruning the space of ITG alignments (Zhang and Gildea, 2006; Haghighi et al., 2009). Pruning in this way reduces the search space dramatically, but only rarely prohibits correct alignments. The oracle alignment error rate for the block ITG model class is 1.4%; the oracle alignment error rate for this pruned subset of ITG is 2.0%.

To take advantage of the sparsity that results from pruning, we use an agenda-based parser that orders search states from small to large, where we define the size of a bispan as the total number of words contained within it. For each size, we maintain a separate agenda. Only when the agenda for size k is exhausted does the parser proceed to process the agenda for size $k+1$.

We also employ coarse-to-fine search to speed up inference (Charniak and Caraballo, 1998). In the coarse pass, we search over the space of ITG alignments, but score only features on alignment links and bispans that are local to terminal blocks. This simplification eliminates the need to augment grammar symbols, and so we can exhaustively explore the (pruned) space. We then compute outside scores for bispans under a max-sum semiring (Goodman, 1996). In the fine pass with the full extraction set model, we impose a maximum size of 10,000 for each agenda. We order states on agendas by the sum of their inside score under the full model and the outside score computed in the coarse pass, pruning all states not within the fixed agenda beam size.

Search states that are popped off agendas are indexed by their corner locations for fast look-up when constructing new states. For each corner and size combination, built states are maintained in sorted order according to their inside score. This ordering allows us to stop combining states early when the results are falling off the agenda beams. Similar search and beaming strategies appear in many decoders for machine translation (Huang and Chiang, 2007; Koehn and Haddow, 2009; Moore and Quirk, 2007).

4.3 Finding Pseudo-Gold ITG Alignments

Equation 3 asks for the block ITG alignment $\mathcal{A}_g$ that is closest to a reference alignment $\mathcal{A}_t$, which may not lie in ITG(**e**,**f**). We search for

[6] The number of configuration states does not depend on the size of $\mathcal{A}$ because corners have fixed size, and because the position of links within rows or columns is not needed.

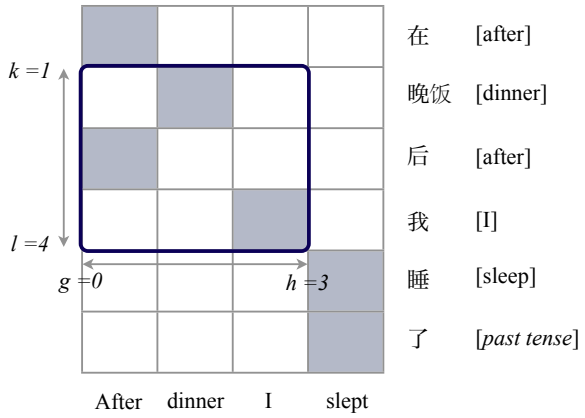

Figure 7: A* search for pseudo-gold ITG alignments uses an admissible heuristic for bispans that counts the number of gold links outside of $[k, \ell)$ but within $[g, h)$. Above, the heuristic is 1, which is also the minimal number of alignment errors that an ITG alignment will incur using this bispan.

$\mathcal{A}_g$ using A* bitext parsing (Klein and Manning, 2003). Search states, which correspond to bispans $[g, h) \Leftrightarrow [k, \ell)$, are scored by the number of errors within the bispan plus the number of $(i, j) \in A_t$ such that $j \in [k, \ell)$ but $i \notin [g, h)$ (recall errors). As an admissible heuristic for the future cost of a bispan $[g, h) \Leftrightarrow [k, \ell)$, we count the number of $(i, j) \in \mathcal{A}_t$ such that $i \in [g, h)$ but $j \notin [k, \ell)$, as depicted in Figure 7. These links will become recall errors eventually. A* search with this heuristic makes no errors, and the time required to compute pseudo-gold alignments is negligible.

5 Relationship to Previous Work

Our model is certainly not the first alignment approach to include structures larger than words. Model-based phrase-to-phrase alignment was proposed early in the history of phrase-based translation as a method for training translation models (Marcu and Wong, 2002). A variety of unsupervised models refined this initial work with priors (DeNero et al., 2008; Blunsom et al., 2009) and inference constraints (DeNero et al., 2006; Birch et al., 2006; Cherry and Lin, 2007; Zhang et al., 2008). These models fundamentally differ from ours in that they stipulate a segmentation of the sentence pair into phrases, and only align the minimal phrases in that segmentation. Our model scores the larger overlapping phrases that result from composing these minimal phrases.

Discriminative alignment is also a well-explored area. Most work has focused on predicting word alignments via partial matching inference algorithms (Melamed, 2000; Taskar et al., 2005; Moore, 2005; Lacoste-Julien et al., 2006). Work in semi-supervised estimation has also contributed evidence that hand-annotations are useful for training alignment models (Fraser and Marcu, 2006; Fraser and Marcu, 2007). The ITG grammar formalism, the corresponding word alignment class, and inference procedures for the class have also been explored extensively (Wu, 1997; Zhang and Gildea, 2005; Cherry and Lin, 2007; Zhang et al., 2008). At the intersection of these lines of work, discriminative ITG models have also been proposed, including one-to-one alignment models (Cherry and Lin, 2006) and block models (Haghighi et al., 2009). Our model directly extends this research agenda with first-class possible links, overlapping phrasal rule features, and an extraction-level loss function.

Kääriäinen (2009) trains a *translation* model discriminatively using features on overlapping phrase pairs. That work differs from ours in that it uses fixed word alignments and focuses on translation model estimation, while we focus on alignment and translate using standard relative frequency estimators.

Deng and Zhou (2009) present an alignment combination technique that uses phrasal features. Our approach differs in two ways. First, their approach is tightly coupled to the input alignments, while we perform a full search over the space of ITG alignments. Also, their approach uses greedy search, while our search is optimal aside from pruning and beaming. Despite these differences, their strong results reinforce our claim that phrase-level information is useful for alignment.

6 Experiments

We evaluate our extraction set model by the bispans it predicts, the word alignments it generates, and the translations generated by two end-to-end systems. Table 1 compares the five systems described below, including three baselines. All supervised aligners were optimized for bispan F_5.

Unsupervised Baseline: GIZA++. We trained GIZA++ (Och and Ney, 2003) using the default parameters included with the Moses training script (Koehn et al., 2007). The designated regimen concludes by Viterbi aligning under Model 4 in both directions. We combined these alignments with

the *grow-diag* heuristic (Koehn et al., 2003).

Unsupervised Baseline: Joint HMM. We trained and combined two HMM alignment models (Ney and Vogel, 1996) using the Berkeley Aligner.[7] We initialized the HMM model parameters with jointly trained Model 1 parameters (Liang et al., 2006), combined word-to-word posteriors by averaging (soft union), and decoded with the competitive thresholding heuristic of DeNero and Klein (2007), yielding a state-of-the-art unsupervised baseline.

Supervised Baseline: Block ITG. We discriminatively trained a block ITG aligner with only sure links, using block terminal productions up to 3 words by 3 words in size. This supervised baseline is a reimplementation of the MIRA-trained model of Haghighi et al. (2009). We use the same features and parser implementation for this model as we do for our extraction set model to ensure a clean comparison. To remain within the alignment class, MIRA updates this model toward a pseudo-gold alignment with only sure links. This model does not score any overlapping bispans.

Extraction Set Coarse Pass. We add possible links to the output of the block ITG model by adding the mixed terminal block productions described in Section 2.3. This model scores overlapping phrasal rules contained within terminal blocks that result from including or excluding possible links. However, this model does not score bispans that cross bracketing of ITG derivations.

Full Extraction Set Model. Our full model includes possible links and features on extraction sets for phrasal bispans with a maximum size of 3. Model inference is performed using the coarse-to-fine scheme described in Section 4.2.

6.1 Data

In this paper, we focus exclusively on Chinese-to-English translation. We performed our discriminative training and alignment evaluations using a hand-aligned portion of the NIST MT02 test set, which consists of 150 training and 191 test sentences (Ayan and Dorr, 2006). We trained the baseline HMM on 11.3 million words of FBIS newswire data, a comparable dataset to those used in previous alignment evaluations on our test set (DeNero and Klein, 2007; Haghighi et al., 2009).

[7] http://code.google.com/p/berkeleyaligner

Our end-to-end translation experiments were tuned and evaluated on sentences up to length 40 from the NIST MT04 and MT05 test sets. For these experiments, we trained on a 22.1 million word parallel corpus consisting of sentences up to length 40 of newswire data from the GALE program, subsampled from a larger data set to promote overlap with the tune and test sets. This corpus also includes a bilingual dictionary. To improve performance, we retrained our aligner on a retokenized version of the hand-annotated data to match the tokenization of our corpus.[8] We trained a language model with Kneser-Ney smoothing on 262 million words of newswire using SRILM (Stolcke, 2002).

6.2 Word and Phrase Alignment

The first panel of Table 1 gives a word-level evaluation of all five aligners. We use the alignment error rate (AER) measure: precision is the fraction of sure links in the system output that are sure or possible in the reference, and recall is the fraction of sure links in the reference that the system outputs as sure. For this evaluation, possible links produced by our extraction set models are ignored. The full extraction set model performs the best by a small margin, although it was not tuned for word alignment.

The second panel gives a phrasal rule-level evaluation, which measures the degree to which these aligners matched the extraction sets of hand-annotated alignments, $R_3(\mathcal{A}_t)$.[9] To compete fairly, all models were evaluated on the full extraction sets induced by the word alignments they predicted. Again, the extraction set model outperformed the baselines, particularly on the F_5 measure for which these systems were trained.

Our coarse pass extraction set model performed nearly as well as the full model. We believe these models perform similarly for two reasons. First, most of the information needed to predict an extraction set can be inferred from word links and phrasal rules contained within ITG terminal productions. Second, the coarse-to-fine inference may be constraining the full phrasal model to predict similar output to the coarse model. This similarity persists in translation experiments.

[8] All alignment results are reported under the annotated data set's original tokenization.

[9] While pseudo-gold approximations to the annotation were used for training, the evaluation is always performed relative to the original human annotation.

		Word			Bispan				BLEU	
		Pr	Rc	AER	Pr	Rc	F_1	F_5	Joshua	Moses
Baseline models	GIZA++	72.5	71.8	27.8	69.4	45.4	54.9	46.0	33.8	32.6
	Joint HMM	84.0	76.9	19.6	69.5	59.5	64.1	59.9	34.5	33.2
	Block ITG	83.4	83.8	16.4	**75.8**	62.3	68.4	62.8	34.7	33.6
Extraction set models	Coarse Pass	82.2	**84.2**	16.9	70.0	72.9	71.4	72.8	35.7	34.2
	Full Model	**84.7**	84.0	**15.6**	69.0	**74.2**	**71.6**	**74.0**	**35.9**	**34.4**

Table 1: Experimental results demonstrate that the full extraction set model outperforms supervised and unsupervised baselines in evaluations of word alignment quality, extraction set quality, and translation. In *word* and *bispan* evaluations, GIZA++ did not have access to a dictionary while all other methods did. In the *BLEU* evaluation, all systems used a bilingual dictionary included in the training corpus. The *BLEU* evaluation of supervised systems also included rule counts from the Joint HMM to compensate for parse failures.

6.3 Translation Experiments

We evaluate the alignments predicted by our model using two publicly available, open-source, state-of-the-art translation systems. Moses is a phrase-based system with lexicalized reordering (Koehn et al., 2007). Joshua (Li et al., 2009) is an implementation of Hiero (Chiang, 2007) using a suffix-array-based grammar extraction approach (Lopez, 2007).

Both of these systems take word alignments as input, and neither of these systems accepts possible links in the alignments they consume. To interface with our extraction set models, we produced three sets of sure-only alignments from our model predictions: one that omitted possible links, one that converted all possible links to sure links, and one that includes each possible link with 0.5 probability. These three sets were aggregated and rules were extracted from all three.

The training set we used for MT experiments is quite heterogenous and noisy compared to our alignment test sets, and the supervised aligners did not handle certain sentence pairs in our parallel corpus well. In some cases, pruning based on consistency with the HMM caused parse failures, which in turn caused training sentences to be skipped. To account for these issues, we added counts of phrasal rules extracted from the baseline HMM to the counts produced by supervised aligners.

In Moses, our extraction set model predicts the set of phrases extracted by the system, and so the estimation techniques for the alignment model and translation model both share a common underlying representation: extraction sets. Empirically, we observe a BLEU score improvement of 1.2 over the best unsupervised baseline and 0.8 over the block ITG supervised baseline (Papineni et al., 2002).

In Joshua, hierarchical rule extraction is based upon phrasal rule extraction, but abstracts away sub-phrases to create a grammar. Hence, the extraction sets we predict are closely linked to the representation that this system uses to translate. The extraction model again outperformed both unsupervised and supervised baselines, by 1.4 BLEU and 1.2 BLEU respectively.

7 Conclusion

Our extraction set model serves to coordinate the alignment and translation model components of a statistical translation system by unifying their representations. Moreover, our model provides an effective alternative to phrase alignment models that choose a particular phrase segmentation; instead, we predict many overlapping phrases, both large and small, that are mutually consistent. In future work, we look forward to developing extraction set models for richer formalisms, including hierarchical grammars.

Acknowledgments

This project is funded in part by BBN under DARPA contract HR0011-06-C-0022 and by the NSF under grant 0643742. We thank the anonymous reviewers for their helpful comments.

References

Necip Fazil Ayan and Bonnie J. Dorr. 2006. Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proceedings of*

the Annual Conference of the Association for Computational Linguistics.

Necip Fazil Ayan, Bonnie J. Dorr, and Christof Monz. 2005. Neuralign: combining word alignments using neural networks. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Alexandra Birch, Chris Callison-Burch, and Miles Osborne. 2006. Constraining the phrase-based, joint probability statistical translation model. In *Proceedings of the Conference for the Association for Machine Translation in the Americas*.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. 2009. A Gibbs sampler for phrasal synchronous grammar induction. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Eugene Charniak and Sharon Caraballo. 1998. New figures of merit for best-first probabilistic chart parsing. In *Computational Linguistics*.

Colin Cherry and Dekang Lin. 2006. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Colin Cherry and Dekang Lin. 2007. Inversion transduction grammar for joint phrasal translation modeling. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics Workshop on Syntax and Structure in Statistical Translation*.

David Chiang, Yuval Marton, and Philip Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Koby Crammer and Yoram Singer. 2003. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3:951–991.

John DeNero and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

John DeNero and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proceedings of the Annual Conference of the Association for Computational Linguistics: Short Paper Track*.

John DeNero, Dan Gillick, James Zhang, and Dan Klein. 2006. Why generative phrase models underperform surface heuristics. In *Proceedings of the NAACL Workshop on Statistical Machine Translation*.

John DeNero, Alexandre Bouchard-Cote, and Dan Klein. 2008. Sampling alignment structure under a bayesian translation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Yonggang Deng and Bowen Zhou. 2009. Optimizing word alignment combination for phrase table training. In *Proceedings of the Annual Conference of the Association for Computational Linguistics: Short Paper Track*.

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Alexander Fraser and Daniel Marcu. 2007. Getting the structure right for word alignment: Leaf. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Joshua Goodman. 1996. Parsing algorithms and metrics. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. 2009. Better word alignments with supervised ITG models. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Matti Kääriäinen. 2009. Sinuhe—statistical machine translation using a globally trained conditional exponential family translation model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Dan Klein and Chris Manning. 2003. A* parsing: Fast exact Viterbi parse selection. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Philipp Koehn and Barry Haddow. 2009. Edinburghs submission to all tracks of the WMT2009 shared task with reordering and speed improvements to Moses. In *Proceedings of the Workshop on Statistical Machine Translation*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics: Demonstration track*.

Simon Lacoste-Julien, Ben Taskar, Dan Klein, and Michael I. Jordan. 2006. Word alignment via quadratic assignment. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren Thornton, Jonathan Weese, and Omar Zaidan. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *Proceedings of the Workshop on Statistical Machine Translation*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Daniel Marcu and Daniel Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*.

Robert Moore and Chris Quirk. 2007. Faster beam-search decoding for phrasal statistical machine translation. In *Proceedings of MT Summit XI*.

Robert C. Moore. 2005. A discriminative framework for bilingual word alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hermann Ney and Stephan Vogel. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the Conference on Computational linguistics*.

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

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved alignment models for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Andreas Stolcke. 2002. Srilm an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*.

Ben Taskar, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

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

Hao Zhang and Daniel Gildea. 2005. Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.

Hao Zhang and Daniel Gildea. 2006. Efficient search for inversion transduction grammar. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Hao Zhang, Chris Quirk, Robert C. Moore, and Daniel Gildea. 2008. Bayesian learning of non-compositional phrases with synchronous parsing. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*.