Towards Syntactically Contrained Statistical Word Alignment

Greg Hanneman
11-734: Advanced Machine Translation Seminar
White Paper Report

May 5, 2008

1 Introduction

In statistical machine translation, the fundamental problem of word alignment is the process of finding word-to-word connections (i.e. translations) across languages given a sentence in one language and its translation in another. In more formal terms, given a source-language sentence F of n words $(f_1, f_2, ..., f_n)$ and a target-language sentence E of m words $(e_1, e_2, ..., e_m)$, an alignment is a mapping between subsets of F (elements of the power set 2^F) and subsets of E (elements of 2^E). Instead of a mapping between full subsets, an alignment is usually indicated as a collection of links, each of which connects some f_j $(1 \leq j \leq n)$ to some e_i $(1 \leq i \leq m)$. The total collection of links makes up the alignment for the given sentence pair.

In the general case, the total number of possible alignments, called the alignment space, is extremely large. With no restrictions in place and an n-word sentence pair, there are n^2 possible alignment links and 2^{n^2} possible alignments. If a one-to-one constraint is enforced, such that one word in F may only align to one word in E, this exponential space can be reduced to $n!$. Additional constraints may further restrict the alignment space or lead to related spaces (Cherry and Lin, 2006a).

The natural goal of constrained alignment is to restrict the alignment space in such a way that "bad" or linguistically very unlikely alignments are ruled out while "good" or linguistically sound alignments remain possible or are preferred. Word alignment is most commonly carried out within the scope of a parallel sentence represented as a flat stream of plain-text words or as a flat stream of sets of feature–value pairs. However, in the realm of natural language, it is also possible to represent the structure inherent in a sentence; further, the structure can provide useful information about what alignments are "good" and what alignments are "bad" beyond what information can be extracted from a flat string. In this paper, we will consider a number of techniques for representing different levels of syntactic structure in the alignment process and examine the benefit of the information it provides.

First, in Section 2, we briefly describe basic statistical alignment models that do not take into account any overt representation of the syntax of the sentence they are aligning. A number of published extensions to or replacements for the base models will be discussed in Section 3; these approaches all explicitly model some level of structure on one or both sides of the parallel sentence pair. Section 4 considers tradeoffs that these models introduce, compares their expressive and restrictive powers, and concludes the paper with some possible avenues for future alignment research.

2 Non-Syntax-Based Methods

Since the introduction of statistical machine translation in 1990, the most common approach to word alignment as a subtask has been with generative models. In the past few years, however, work has been carried out in discriminative approaches as well. In this section, we will review the IBM models (the standard generative approach) and a discriminative alignment technique that has shown promising results.

2.1 Generative Models

A generative approach to word alignment begins with a generative story: for an alignment from English to French, for example, the story is that each of the words in the English sentence individually generated a certain number of the words in the French sentence (Brown et al., 1990); the task of alignment is to recover these correspondences. The generation of the French sentence occurs in a series of three steps. In an initial fertility step, each English word produces a certain number of French word placeholders. A lexical production step fills in French words for each of these placeholders, and then the words may be reordered in a final distortion step. Together, the fertility, lexical translation, and distortion models specify the probability of producing the French sentence given the initial English sentence: $P(F \,|\, E)$.

The *de facto* models of statistical word alignment in recent years have been those developed by IBM in 1993 (Brown et al., 1993) and implemented more recently in the GIZA++ toolkit (Och and Ney, 2003). Under these generative formulations, the alignment of a sentence is modeled as a hidden variable; instead of computing the probability of a source sentence given the target, $P(F \,|\, E)$, the calculation also takes into account the alignment of the sentence pair, $P(F, A \,|\, E)$. In this case, an element a_j of $A = a_1, a_2, ..., a_j$ represents the word index (or indexes, more generally) in E that f_j is aligned to.

The five original IBM models range in complexity, statistical deficiency, and amount of linguistic knowledge encapsulated, although none of the models takes any explicit account of the syntax of the sentences being aligned. Model 1 performs one-to-one alignments and treats F and E as simple "bags of words," meaning that the word order plays no role in the sentence alignment:

$$P(F, A \,|\, E) = \frac{P(n \,|\, m)}{(m+1)^n} \prod_{j=1}^{n} P(f_j \,|\, e_{a_j}) \tag{1}$$

Here, the first term on the right-hand side provides a probability distribution over the length of the sentence pair, and $P(f_j \,|\, e_{a_j})$ is the lexical probability of f_j given the word in E it is aligned to. In Model 2, the probability of an alignment link depends on the positions of the words being aligned:

$$P(F, A \,|\, E) = P(n \,|\, m) \prod_{j=1}^{n} \Big(P(a_j \,|\, j, n, m) P(f_j \,|\, e_{a_j}) \Big) \tag{2}$$

Thus, Model 1 is just a special case of Model 2 where the alignment-position probability $P(a_j \,|\, j, n, m)$ is just a uniform $\frac{1}{m+1}$.

Later IBM models begin to include more linguistic intuition with the introduction of fertility and distortion models. The concept of fertility represents the linguistic observation that some source words may more naturally translate into the target language as multiple words (English *not* to French *ne ... pas*, for example) or as no words at all. Model 3 expresses this as $P(\phi \,|\, e_i)$, the probability that a word e_i will generate ϕ words in F. Distortion, which also appears beginning in Model 3, considers permutations of the words in F and assigns probabilities to them. Beginning in Model 4, a first-order dependency is introduced into the alignment model that specifies a distribution for a fertile alignment (e_i, $f_{j1}\ f_{j2}\ ...\ f_{jk}$) over the placements of f_{j2} through f_{jk}.

In addition to the IBM models, another commom discriminative approach to word alignment is the HMM model of Vogel, Ney, and Tillmann (1996), which also includes a first-order dependency by means of a hidden Markov model. It represents the linguistic intuition that words in parallel sentences often tend to group together into clusters by conditioning the alignment for f_j on the difference in position between it and the alignment for f_{j-1}:

$$P(F, A \,|\, E) = \prod_{j=1}^{n} \Big(P(a_j \,|\, a_{j-1}, m) P(f_j \,|\, e_{a_j}) \Big) \tag{3}$$

2.2 Discriminative Models

In contrast to the generative approach described above, work has also been carried out on taking a discriminative view of the word alignment problem. In the discriminative approach, all possible alignments are considered and assigned scores; the final output is the matching with the highest total score, possibly subject to some set of constraints. The discriminative framework is well suited for incorporating arbitrary features to make up the score for each alignment link, which is more difficult to do in carefully factored statistical generative models. On the other hand, discriminative training relies on at least a small amount of gold-standard training data, whereas generative models are completely unsupervised.

Taskar, Lacoste-Julien, and Klein (2005) proposed a general, but non-syntax-aware, discriminative method for computing scores of possible alignments based on a weighted function of arbitrary features — including, if desired, predictions from the IBM generative models. Each possible alignment link (e_i, f_j) is assigned a score $v(e_i, f_j) = \mathbf{w}^T \mathbf{f}(e_i, f_j)$, where $\mathbf{f}(e_i, f_j)$ is a vector of feature values for the link and $\mathbf{w}$ is a vector of learned weights for those features. The authors model a number of features on the type of information represented in the IBM models, with the added constraint that all of their alignments are either one-to-one or one-to-zero. These include co-occurrence features (Dice coefficients on pairs of words), position difference features, and co-occurrence features for the words following the current pair being aligned (to approximate first-order dependency features from the IBM and HMM models). In addition, there are word-string similarity features and particular lexical translation features for very high-frequency words. IBM Model 4 predictions can also be added as another feature; when they are included, Taskar, Lacoste-Julien, and Klein (2005) report the lowest published alignment error rate (AER) for a French–English alignment task.

3 General Syntax-Based Approaches

In extending — or replacing — the general approaches to word alignment discussed in Section 2, syntactic information has the potental to provide useful guidance during the search of an alignment space to rule out or de-emphasize incorrect alignments and shift greater likelihood to correct alignments. The movement of a multi-word noun phrase in translation from an SOV to a VSO language, for example, can be modeled as a single operation on the phrase rather than a series of (at least somewhat) independent operations on individual words (Cherry and Lin, 2006a). This type of modeling also has the simultaneous effect of restricting the alignment space in a meaningful way: if a noun phrase e_1, e_2, e_3 must be aligned as one unit, as in the previous example, syntactically and intuitively incorrect alignments for it such as f_1, f_7, f_8 are immediately ruled out. The enforcement of such a phrasal cohesion constraint is usually maintained in gold-standard alignments (Cherry and Lin, 2006b).

Further, DeNero and Klein (2007) provide a motivational example for why syntax-based word alignment can be important for syntax-aware translation systems. In their example (Figure 1), incorrect word alignments make it impossible for further steps in a syntax-based MT system to extract transfer rules from given sentence pairs. Given the word alignments in Figure 1, a baseline phrase extractor succeeds at extracting the phrase pair (*jobs are*, *emplois sont*). However, the incorrect alignment between *the* and *la* makes it impossible for a syntax-based phrase extractor to extract any contiguous French spans corresponding to constituents in the English parse tree.

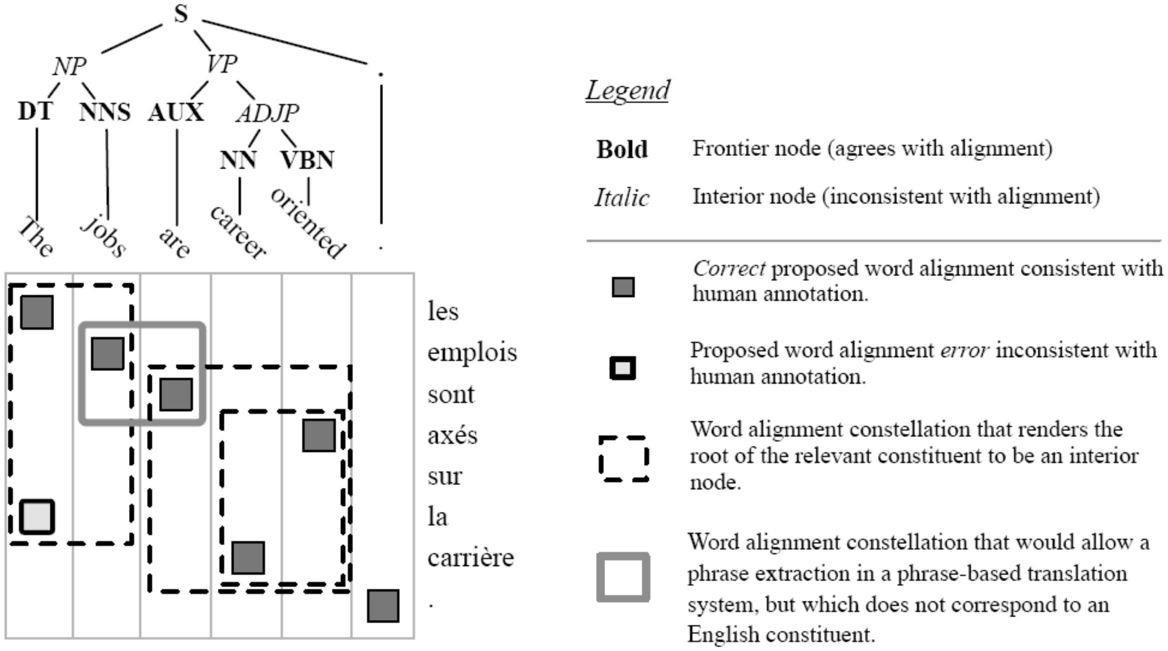
S
NP
VP
DT
NNS
AUX
ADJP
NN
VBN
The
jobs
are
career
oriented
.
les
emplois
sont
axés
sur
la
carrière
.
Legend
Bold
Frontier node (agrees with alignment)
Italic
Interior node (inconsistent with alignment)
Correct proposed word alignment consistent with human annotation.
Proposed word alignment error inconsistent with human annotation.
Word alignment constellation that renders the root of the relevant constituent to be an interior node.
Word alignment constellation that would allow a phrase extraction in a phrase-based translation system, but which does not correspond to an English constituent.

Figure 1: Word alignments that are incompatible with a target-language parse may severely impact syntax-based MT. Here, the incorrect (*the*, *la*) alignment prevents any constituent phrase from being extracted (DeNero and Klein, 2007).

3.1 Syntax-Based Distortion Model

To address the above problem, DeNero and Klein (2007) modify the distortion component of the basic HMM model (Vogel, Ney, and Tillmann, 1996). Instead of basing the alignment probability for a_j on the difference in string position between a_j and a_{j-1}, the new syntax-based distortion component models the likelihood of a_j given the path in a parse tree between a_{j-1} and a_j, represented as a series of "pop" operations moving from child to parent in the tree, a "move" operation between two siblings of the same parent, and a series of "push" operations moving from parent to child. Probabilities for these transitions, conditioned on the current node and its parent, siblings, or children, were learned from 100,000 parsed sentences.

On test sets for French–English and Chinese–English alignment, the syntactic-distortion HMM model performed about equally as well as the basic HMM model when evaluated against manually aligned data (Figure 2). Syntax-based distortion led to an improvement in precision in both language pairs, but with a consequent loss in recall. Both HMM models outperformed a GIZA++ training of the IBM models on the same data set. Some of the results may be explained, however, by the authors' use of thresholding and combination methods in producing the final alignments: the HMM models were trained in both the E-to-F and F-to-E directions, then combined with a variety of thresholding heuristics and intersection or union types. On the French–English task, the authors report an AER of 0.084 for the syntactic HMM using a hard intersection and no thresholding; this is more comparable to GIZA++'s AER of 0.086 under the same conditions.

	Chinese–English			**French–English**		
Model	**Prec**	**Rec**	**AER**	**Prec**	**Rec**	**AER**
Basic HMM	0.816	0.788	0.198	0.939	0.930	0.065
Syntactic HMM	0.822	0.768	0.205	0.952	0.915	0.064
GIZA++	0.619	0.826	0.297	0.960	0.861	0.086

Figure 2: A modified HMM distortion model (DeNero and Klein, 2007) for Chinese–English and French–English alignment performs comparably to the basic HMM model.

3.2 Alignments from Tree-to-String Models

The tree-to-string model of translation (Yamada and Knight, 2001) modifies the generative story from Section 2.1 to start with a target-language parse tree and transform it into a source-language string; tracing the effects of the transformations specifies a word alignment. In the tree-to-string model, the word-level fertility and distortion steps are replaced with two tree operations: insertion of new lexical-level nodes and reordering of the daughter nodes of a single parent. After the reordering, node insertion, and lexical translation steps, the leaf nodes are read off to produce the output string.

Yamada and Knight (2001) conducted a small initial alignment experiment using the model, which compared its performance to IBM Model 5 on a corpus of 2121 short Japanese–English sentences with English parse trees. Alignments produced on the first 50 sentences of the training set were evaluated by humans as either "OK" (1 point), "not sure" (0.5 points), or "wrong" (0 points). The tree-to-string model received an average score of 0.582 per alignment, while Model 5 scored 0.431. The perplexity of the tree-to-string model on the training set (15.79) was between Model 1 (24.01) and Model 5 (9.84).

The basic tree-to-string model was extended by Gildea (2003) to allow for syntactially well-formed departures from the structure of the target-language parse, thus making the model better equipped to handle divergent linguistic structures or free translations between the source and target sentences. His "loosely tree-based" model introduces the concept of subtree cloning, which permits more expressive reordering than the base tree-to-string model by copying a subtree and inserting it as a new node somewhere else in the parse tree. The operation is governed by two probabilities: for a node n_p, a single parameter controls the probabilty of inserting a cloned subtree as a new child of n_p; then, the root of the subtree to clone is chosen uniformly from any other node n_c in the tree.

Training on a corpus of 4982 parallel Korean–English sentences with manually produced Korean parse trees, and testing on 101 sentences, Gildea (2003) compared the tree-to-string model with subtree cloning to the basic tree-to-string model of Yamada and Knight (2001) and to IBM Models 1, 2, and 3. The results in Figure 3 show a reduction in AER once the subtree cloning operation is introduced, plus a further reduction when the probability of inserting a new lexical-level node to the left of an existing node — one of Yamada and Knight's original model parameters — is manually fixed at 0.5 rather than being estimated during the EM training of the model.

Model	**AER**
IBM Model 1	0.37
IBM Model 2	0.35
IBM Model 3	0.43
Tree-to-String	0.42
Tree-to-String + Cloning	0.36
Tree-to-String + Cloning + Fixed P_{ins}(left)	0.32

Figure 3: Korean–English alignments results show that extending the basic tree-to-string model with a subtree cloning operation (Gildea, 2003) improves alignment error rate.

3.3 Inversion Transduction Grammars

Wu's (1997) Inversion Transduction Grammar (ITG) was originally developed as a formalism for synchronously parsing bilingual text, and thus it provides a word alignment as a side effect. It parses and produces output simultaneously in both a source-side and a target-side output stream. In linguistic terms, the grammar is minimal, representing only a single non-terminal and three context-free grammar rules:

$$A \rightarrow [A\ A] \tag{4}$$

$$A \rightarrow \langle A\ A \rangle \tag{5}$$

$$A \rightarrow f/e \tag{6}$$

In the standard ITG notation, the square brackets in Equation 4 indicate that the right-hand side constituents are produced in the same left-to-right order in both source and target streams; the angled brackets in Equation 5 indicate that the order of the constituents is reversed in the target stream. Equation 6 indicates that a terminal string f is produced in the source stream while e is produced in the target stream.

The simplicity of the ITG formalism has made it a useful starting point for syntax-constrained word alignment under a tree-to-tree model where parse information is taken into account in both the source and target languages. By introducing the constraint that all alignments must be represented as a binary tree with inversions, but not a restrictive grammar, it searches a comparatively large subset of the $n!$ space of one-to-one alignments (Cherry and Lin, 2006a) — exploring almost all of it for short-distance reordering, but rapidly pruning as sentence length gets longer (Wu, 1997). A further optimization can be made by using a tail-recursive "canonical form" ITG that removes redundancy in the search space by deriving only one parse structure for a given alignment (Zhang and Gildea, 2004; Cherry and Lin, 2006a):

$$S \rightarrow A \mid B \mid C \tag{7}$$

$$A \rightarrow [A\ B] \mid [B\ B] \mid [C\ B] \mid [A\ C] \mid [B\ C] \mid [C\ C] \tag{8}$$

$$B \rightarrow \langle A\ A\rangle \mid \langle B\ A\rangle \mid \langle C\ A\rangle \mid \langle A\ C\rangle \mid \langle B\ C\rangle \mid \langle C\ C\rangle \tag{9}$$

$$C \rightarrow f/e \tag{10}$$

Zhang and Gildea (2004) trained a canonical-form ITG parser on 18,773 sentence pairs of parallel Chinese–English data, restricting the set to sentences shorter than 25 words in both languages, and evaluated alignment quality on a test set of 48 sentence pairs of hand-aligned data. For comparison, they also trained IBM Models 1 and 4, along with Yamada and Knight's (2001) basic tree-to-string model and Gildea's (2003) subtree cloning extension to it. The results, shown in Figure 4, indicate that the ITG model has the highest precision, highest recall, and lowest AER even though it is limited to one-to-one alignments. Similar results were shown for a French–English task, using a training set of 20,000 sentence pairs and a larger test set of 447 sentence pairs, although in the French case the ITG model was closely matched by IBM Model 4 and the tree-to-string model with subtree cloning.

	Chinese–English			**French–English**		
Model	**Prec**	**Rec**	**AER**	**Prec**	**Rec**	**AER**
IBM Model 1	0.56	0.42	0.52	0.63	0.71	0.34
IBM Model 4	0.67	0.43	0.47	0.83	0.83	0.17
ITG	0.68	0.52	0.40	0.82	0.87	0.16
Tree-to-String	0.63	0.41	0.50	—	—	—
Tree-to-String Cloned	0.65	0.43	0.48	0.84	0.85	0.15

Figure 4: Alignment results from Zhang and Gildea (2004) for various models on Chinese–English and French–English. The model based on Inversion Transduction Grammar has the best overall performance.

3.4 Dependency-Augmented ITGs

One criticism of using ITG-based constraints on word alignment is that ITGs are linguistically rather arbitrary. The canonical form grammar in Equations 7 though 10 has only four non-terminals, none of which exactly corresponds to a linguistically motivated unit such as a noun phrase, prepositional phrase, adjective, etc. While ITG-constrained alignment does maintain the phrasal cohesion constraint introduced at the beginning of Section 3, it is not guaranteed to do so in a linguistically meaningful way.

ITG efficiency and linguistic phrases can be brought together in a dependency-constrained ITG, whereby the binary-branching constraint from a basic ITG grammar is extended to disallow ITG phrases that span across a constituent boundary from the dependency parse (Cherry and Lin, 2006b). In an ITG chart parser, the effect is accomplished by seeding the invalid spans with a score of either $-\infty$ or a penalty of some intermediate strength. Invalid spans are extractable from a dependency representation as shown in Figure 5: valid spans are those where each head–modifier chain is either completely included or completely excluded.

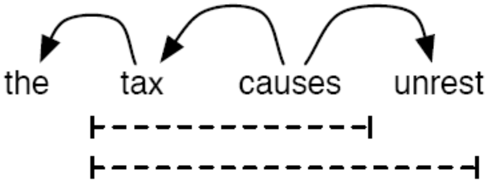

Figure 5: Invalid spans in a dependency tree contain only parts of a head–modifier chain (Cherry and Lin, 2006b).

Cherry and Lin (2006b) experimented with a soft dependency constraint in ITG-driven alignment. Using a support vector machine (SVM), their discriminative training approach builds on the non-syntactic approach of Taskar, Lacoste-Julien, and Klein (2005) described in Section 2.2. With the original features mostly unchanged (Model 4 predictions are excluded, and the function for calculating word-level correlations between e_i and f_j is modified), Cherry and Lin add features for marking inverted ITG rules (those of the type in Equation 5) and invalid spans according to the dependency tree. Features are defined on instances of ITG production rules and summed over the complete parse for a sentence. As discriminative training requires some amount of labelled training examples, the authors created 100 ITG parse trees from gold-standard alignment data by inducing trees that maximize the number of gold-standard alignments while minimizing the number of invalid dependency spans used.

The authors conducted two experiments on French–English alignment. The first tested a hard dependency-constrained ITG parser against an unconstrained ITG and a completely unconstrained discriminative aligner. For each alignment technique, link scores were calculated as

$$v(e_i, f_j) = \phi^2(e_i, f_j) - 10^{-5} \left| \frac{i}{m} - \frac{j}{n} \right| \tag{11}$$

where ϕ^2 is a co-occurrence measure of the words in the link; this allowed the alignment results to be compared without taking supervised learning into account. The results of this warm-up experiment are shown in Figure 6(a). In the second experiment, full supervised learning over feature sets on 100 sentence pairs was included, and hard and soft dependency constraints were compared to the Taskar, Lacoste-Julien, and Klein (2005) non-syntactic baseline. These results are in Figure 6(b).

Figure 6(a) shows that ruling out invalid dependency spans has a beneficial effect on the word alignment, with a 34 percent relative reduction in AER compared to unconstrained alignment. With full feature-based learning turned on, however, Figure 6(b) shows only a 9 percent relative reduction using hard constraints (and a loss in recall), although learning with soft constraints still achieves a 22 percent relative AER reduction (and a better recall).

4 Summary and Discussion

The syntax-based and non-syntax-based alignment models we have considered in this paper highlight a number of tradeoffs in different approaches to the statistical word alignment problem. Many

Model	Prec	Rec	AER
Unconstrained	0.723	0.845	0.231
ITG	0.764	0.860	0.200
ITG + Hard dep.	0.830	0.873	0.153

(a)

Model	Prec	Rec	AER
Unconstrained	0.916	0.860	0.110
ITG + Hard dep.	0.940	0.854	0.100
ITG + Soft dep.	0.944	0.878	0.086

(b)

Figure 6: *(a)* French–English discriminative alignment results for the simple alignment link score in Equation 11, and *(b)* French–English discriminative alignment results with full supervised learning (Cherry and Lin, 2006b).

of the approaches highlight a fundamental decision among levels of model complexity, computability, and statistical correctness. Parameterizations and assumptions made in IBM Models 3 and 4, for example, lead to probability distributions that do not sum up to 1 (Och and Ney, 2003); Model 5, which provides a statistically non-deficient version of Model 4, requires many more parameters. As we noted in Section 2.2, the complicated statistics behind some of the generative models makes them more difficult to adapt, extend, or even compute quickly. Taskar, Lacoste-Julien, and Klein (2005) report that running GIZA++ training through IBM Model 4 took 18 hours to align 1.1 million words, while their own discriminative approach learned feature weights in six minutes on 100 training sentences or in three hours on 5000 training sentences.

Model complexity or model constraints can also come into conflict with the range of linguistic phenomena being represented. Gildea's (2003) extension to the basic Yamada and Knight (2001) tree-to-string model allowed it to capture a wider range of possible alignments for divergent sentence structures; a tree-to-tree vesion of the same subtree cloning model ran 20 times faster and reduced the theoretical complexity in terms of sentence length n from $O(n^4)$ to $O(n^2)$.

This highlights the problem of choosing to model the "right" level of syntactic information or syntactic constraints. The probability calculations in Gildea's subtree cloning operation are minimal, and the success of the operation seems to depend on a surprisingly uninformed chain of independently selecting a useful position in the tree to insert a clone, finding the proper subtree to copy into that location, and generating meaningful lexical translations at the leaves of the cloned tree while generating null words for the same lexical leaves in their original locations. Still, allowing subtree cloning as a softening of Yamada and Knight's original tree constraint improves alignment results in a number of language pairs. (See Figures 3 and 4.) And, in general, hard syntactic constraints may prove to be too rigid and restrictive for representing divergent sentence strucutres or for free translations of parallel sentences. With a discriminant set of features, Cherry and Lin (2006b) were able to relax a hard dependency tree constraint and achieve better alignment performance when the tree constraint could sometimes be violated.

4.1 Alignment Spaces

A most important tradeoff in syntactically constrained alignment is being able to rule out incorrect alignments while not eliminating correct ones. Cherry and Lin (2006a) investigated this by comparing a number of constrained alignment spaces reflective of the syntactic models we have discussed in this paper:

- Permutation space: One-to-one alignments with reorderings allowed. This is the space searched by IBM Model 1 and the discriminative approach by Taskar, Lacoste-Julien, and Klein (2005); for a sentence of length n, there are $n!$ possible alignments.

- ITG space: Permutation space where reorderings satisfy a binary tree constraint with inversions (Wu, 1997). For short sentences, this is nearly all of permutation space; for longer sentences, the percentage of permutation space covered becomes rapidly smaller.
- Dependency space: Permutation space where phrasal cohesion is maintained. For a dependency tree of one head and $n - 1$ modifiers, this is permutation space; in the other extreme, a dependency tree forming a single chain reduces the alignment space to 2^n.
- D-ITG space: The intersection of dependency space with ITG space. This is the space searched by Cherry and Lin's (2006b) dependency-constrained ITG model.
- HD-ITG space: A subset of D-ITG space where each valid dependency span must also contain a head word. This space is defined by Cherry and Lin (2006a) as an attempt to keep modifiers from grouping together in counterintuitive or non-linguistically motivated constituents.

Two experiments explored the effectiveness of these spaces in terms of representational power. In the first, three searches of permutation space were compared to searches of ITG, dependency, D-ITG, and HD-ITG space, using a lexical co-occurrence score of the form in Equation 11 to evaluate each alignment link. The results showed that more complete searches of permuatation space performed better, as did each successive syntax-based restriction of it. Abridged results are shown in Figure 7(a). The second experiment considered the restrictiveness of each alignment space, or the degree to which it ruled out correct alignments: the link score $v(e_i, f_j)$ was set equal to 1 if the alignment (e_i, f_j) appeared in the gold-standard data, 0 if f_j was null, or -1 if (e_i, f_j) was an incorrect alignment according to the gold standard. The number of gold-standard alignments missed under each alignment space is shown in Figure 7(b).

Model	**AER**
Permutation space	0.192
ITG space	0.174
Dependency space	0.134
D-ITG space	0.133
HD-ITG space	0.132

(a)

Model	**# Missed**
Permutation space	162
ITG space	165
Dependency space	260
D-ITG space	232
HD-ITG space	258

(b)

Figure 7: Comparison of alignment spaces *(a)* using the alignment link score in Equation 11, and *(b)* by number of gold-standard alignments missed (Cherry and Lin, 2006a).

ITG-constrained alignment appears especially advantageous in Figure 7: it rules out almost no correct alignments when compared to a complete search of permutation space, but its constraints achieve a noteworthy reduction in AER. Adding dependency constraints (D-ITG space) also appears quite beneficial to higher alignment quality, although at an increased risk of being unable to produce a larger number of the gold-standard alignments. The other search spaces (permutation, dependency, and HD-ITG) either perform poorly or rule out large numbers of correct alignments for a proportially minimal gain in AER.

4.2 Future Work and Questions

Despite the improved alignment quality that syntactically constrained models can provide, syntax-aware alignment techniques are not yet on equal footing with the more established IBM models in terms of use or usability. One reason for this may be the lack of an "off-the-shelf" training toolkit such as GIZA++ (Och and Ney, 2003), which has become the definitive implementation of the generative IBM and HMM models in recent years and has saved system developers the effort of working out the models and the training schedule from scratch. With a similar toolkit for easily training a discriminative or a syntax-aware alignment model, we may see rapid advances in syntactic alignment technology based on the sheer number of researchers who will be using it.

A number of research questions remain to be worked out. Many of the results reported in this paper were obtained from fairly small training corpora. As an extreme example, Yamada and Knight's (2001) tree-to-string alignment experiment was conducted on just over 2000 short sentences of parallel text. As it is well known that EM-trained generative models often require large amount of data to learn reasonable parameters or to produce reasonable output, the authors' reported IBM Model 5 score is likely to improve drastically with more training data, possibly overtaking the syntactic alignment method in performance. A small (5000-sentence) training corpus may be the key to explaining the performance of the IBM models reported in Figure 3: generally, the more expressive models (such as Model 3) perform better, but their increased performance may require more training data to learn optimal values for an increased number of parameters. Results reported in this paper show that syntactically constrained alignment has been beneficial on relatively small training sets, but the results — both performance and training time or computability, compared to the IBM models — appear to be untested so far on a parallel corpus of, perhaps, 1 million sentence pairs that is more representative of "real-world" MT system development conditions.

A further concern is that the improvements seen on AER may not entail improved end-to-end MT performance. The experiments reported in this paper have been limited to discussions of AER results; there has as yet not been a comparison of MT system output based on different alignment techniques, and some researchers have questioned the correlation between improved AER and improved MT. Future experimentation will need to show that better solutions to the word alignment as a subtask also have a noticable benefit in the overall translation pipeline.

Finally, as time goes by, the range of syntactic features incorporated into the alignment models may become richer. Combining a syntax-aware alignment with an extensible discriminative training framework would allow for the easy incorporation of new and different types of information. In addition to dependency spans or ITG constraints, statistical alignment could make use of part-of-speech information, morphological analyses, or semantic information, for instance.

References

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

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

Cherry, Colin and Dekang Lin. 2006a. A comparison of syntactically motivated alignment spaces. In *Proceedings of EACL 2006*, pages 145–152, Trento, Italy, April.

Cherry, Colin and Dekang Lin. 2006b. Soft syntactic constraints for word alignment through discriminative training. In *Proceedings of the COLING/ACL 2006 Poster Session*, pages 105–112, Sydney, Austrailia, July.

DeNero, John and Dan Klein. 2007. Tailoring word alignments to syntactic machine translation. In *Proceedings of ACL 2007*, pages 17–24, Prague, Czech Republic, June.

Gildea, Daniel. 2003. Loosely tree-based alignment for machine translation. In *Proceedings of ACL 2003*, pages 80–87, Sapporo, Japan, July.

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

Taskar, Ben, Simon Lacoste-Julien, and Dan Klein. 2005. A discriminative matching approach to word alignment. In *Proceedings of HLT/EMNLP 2005*, pages 73–80, Vancouver, Canada, October.

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

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

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

Zhang, Hao and Daniel Gildea. 2004. Syntax-based alignment: Supervised or unsupervised? In *Proceedings of COLING 2004*, pages 418–424, Geneva, Switzerland, August.