Accurate Non-Hierarchical Phrase-Based Translation

Michel Galley and **Christopher D. Manning**
Computer Science Department
Stanford University
Stanford, CA 94305
{mgalley,manning}@cs.stanford.edu

Abstract

A principal weakness of conventional (i.e., non-hierarchical) phrase-based statistical machine translation is that it can only exploit continuous phrases. In this paper, we extend phrase-based decoding to allow both source and target phrasal discontinuities, which provide better generalization on unseen data and yield significant improvements to a standard phrase-based system (Moses). More interestingly, our discontinuous phrase-based system also outperforms a state-of-the-art hierarchical system (Joshua) by a very significant margin (+1.03 BLEU on average on five Chinese-English NIST test sets), even though both Joshua and our system support discontinuous phrases. Since the key difference between these two systems is that ours is not hierarchical—i.e., our system uses a string-based decoder instead of CKY, and it imposes no hard hierarchical reordering constraints during training and decoding—this paper sets out to challenge the commonly held belief that the tree-based parameterization of systems such as Hiero and Joshua is crucial to their good performance against Moses.

1 Introduction

Phrase-based machine translation models (Och and Ney, 2004) advanced the state of the art by extending the basic translation unit from words to phrases. By conditioning translations on more than a single word, a statistical machine translation (SMT) system benefits from the larger context of a phrase pair to properly handle multi-word units and local reorderings. Experimentally, it was found that longer phrases yield better MT output (Koehn et al., 2003). However, while it is computationally feasible at training time to extract phrase pairs of nearly unbounded size (Zhang and Vogel, 2005; Callison-Burch et al., 2005), phrase pairs applicable at test time tend to be fairly short. Indeed, data sparsity often forces conventional phrase-based systems to segment test sentences into small phrases, and therefore to translate dependent words (e.g., the French *ne ... pas*) separately instead of jointly.

We present a solution to this sparsity problem by going beyond using only *continuous phrases*, and instead define our translation unit as any subset of words of a sentence, i.e., a *discontinuous phrase*. We generalize conventional multi-beam string-based decoding (Koehn, 2004) to allow variable-size discontinuities in both source and target phrases. Since each sentence pair can be more flexibly decomposed into translation units, it is possible to exploit the rich context of longer (possibly discontinuous) phrases to improve translation quality. Our decoder provides two extensions to Moses (Koehn et al., 2007): (a) to cope with source gaps, we follow (Lopez, 2007) to efficiently find all discontinuous phrases in the training data that also appear in the input sentence; (b) to enable target discontinuities, we augment translation hypotheses to not only record the current partial translation, but also a set of subphrases that may be appended to the partial translation at some later stages of decoding. With these enhancements, our best discontinuous system outperforms Moses with lexicalized reordering by 0.77 BLEU and 1.53 TER points on average.

We also show that our approach compares favorably to binary synchronous context-free grammar (2-SCFG) systems such as Hiero (Chiang, 2007), even though 2-SCFG systems also allow phrasal discontinuities. Part of this difference may be due to a difference of expressiveness, since 2-SCFG models impose hard hierarchical constraints that our models do not impose. Recent work (Wellington et al., 2006; Søgaard and Kuhn, 2009; Søgaard and

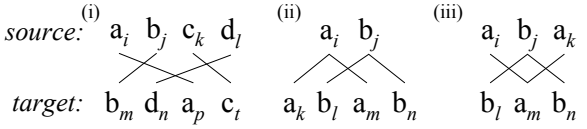


Figure 1: 2-SCFG systems such as Hiero are unable to independently generate translation units **a**, **b**, **c**, and **d** with the following types of alignments: (i) inside-out (Wu, 1997); (ii) cross-serial DTU (Søgaard and Kuhn, 2009); (iii) "bonbon" (Simard et al., 2005). Standard phrase-based decoders cope with (i), but not (ii) and (iii). Our phrase-based decoder handles all three cases.

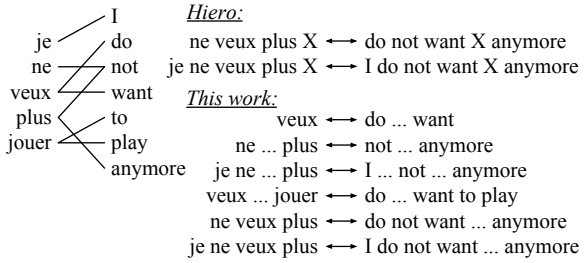


Figure 2: Due to hierarchical constraints, Hiero only extracts two discontinuous phrases from the alignment on the left, but our system extracts 11 (only 6 are shown).

Wu, 2009) has questioned the empirical adequacy of 2-SCFG systems, which are unable to perform any of the transformations shown in Fig. 1. For instance, using manually-aligned bitexts for 12 European languages pairs, Søgaard and Kuhn found that inside-out and cross-serial discontinuous translation units (DTU) account for 1.6% (Danish-English) to 18.6% (French-English) of all translation units. The empirical adequacy of 2-SCFG models would presumably be lower with automatically-aligned texts and if the study also included non-European languages. In contrast, phrase-based systems can properly handle inside-out alignments when used with a reasonably large distortion limit, and all configurations in Fig. 1 are accounted for in our system. In our experiments, we show that our discontinuous phrase-based system outperforms Joshua (Li et al., 2009), a reimplementation of Hiero, by 1.03 BLEU points and 1.19 TER points on average. A final compelling advantage of our decoder is that it preserves the computational efficiency of Moses (i.e., time complexity is linear when a distortion limit is used), while SCFG decoders have a running time that is at least cubic (Huang et al., 2005).

2 Discontinuous Phrase Extraction

In this section, we introduce the extraction of discontinuous phrases for phrase-based MT. We will describe a decoder that can handle such phrases in the next section. Formally, we define the discontinuous phrase-based translation problem as follows. We are given a source sentence $\mathbf{f} = f_1^J = f_1, \ldots, f_j, \ldots, f_J$, which is to be translated into a target sentence $\mathbf{e} = e_1^I = e_1, \ldots, e_i, \ldots, f_I$. Unlike (Och and Ney, 2004), in this work, a sentence pair may be segmented into phrases that are not continuous, so each phrase is characterized by a coverage set, i.e., a set of word indices. Assuming that the sentence pair $(\mathbf{f}, \mathbf{e})$ is decomposed into K discontinuous phrases, we use $\mathbf{s} = (s_1, \ldots, s_K)$ and $\mathbf{t} = (t_1, \ldots, t_K)$ to respectively represent the decomposition of the source and target sentence into K word subsets that are complementary and non-overlapping. A pair of coverage sets (s_k, t_k) is said to be *consistent* with the word alignment A if the following condition holds:

$$\forall (i,j) \in A : i \in s_k \longleftrightarrow j \in t_k \qquad (1)$$

For continuous phrases, finding all phrase pairs that satisfy this condition can be done in $O(nm^3)$ time (Och and Ney, 2004), where n is the length of the sentence and m is the maximum phrase length. The set of discontinuous phrases is exponential in the maximum span length, so phrase extraction must be tailored to a specific text (e.g., a given test sentence) for relatively large m values. Lopez (2007) presents an efficient solution using suffix arrays for finding all discontinuous phrases of the training data that are relevant to a given test sentence or test set. A complete overview of this technique is beyond the scope of this paper, though we will mention that it solves a phrase collocation problem by efficiently identifying collocated continuous phrases of the training data that also happen to be collocated in the test sentence. While this technique was primarily designed for extracting *hierarchical phrases* for Hiero (Chiang, 2007), it can readily be applied to the problem of finding all discontinuous phrases for our phrase-based system. Indeed, the suffix-array technique gives us for each input sentence a list of relevant source coverage sets. For each such s_k, we can easily enumerate each t_k satisfying Eq. 1. The

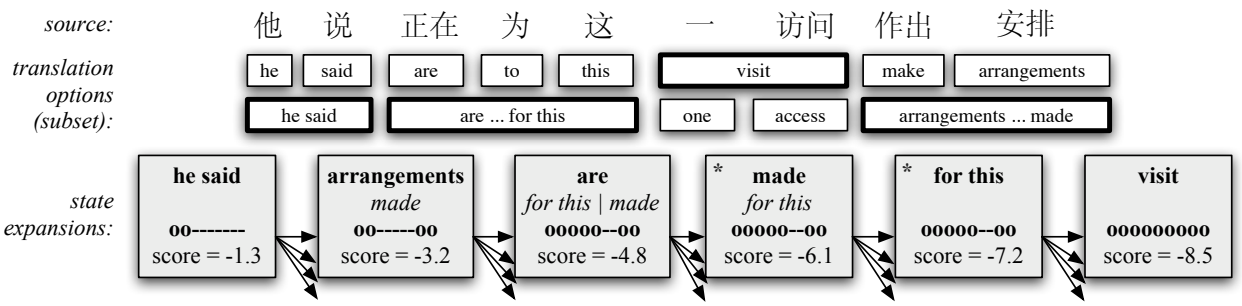


Figure 3: A particular decoder search path for the input shown at the top. Note that this example contains a cross-serial DTU (which interleaves *arrangements ... made* with *are ... for this*), a structure Hiero can't handle.

key difference between Hiero-style extraction and our work is that Eq. 1 is the *only* constraint.[1] Since our decoder doesn't impose hierarchical constraints, we exploit *all* discontinuous phrase pairs consistent with the word alignment, which often includes sound translations not captured by Hiero (e.g., *ne ... plus* translating to *not ... anymore* in Fig. 2).

3 Decoder

The core engine of our phrase-based system, Phrasal (Cer et al., 2010), is a multi-stack decoder similar to Moses (Koehn, 2004), which we extended to support variable-size gaps in the source and the target. In Moses, partial translation hypotheses are arranged into different stacks according to the total number of input words they cover. At every translation step, stacks are pruned using partial translation cost and a lower bound on the estimated future cost. Pruning is implemented using both threshold and histogram pruning, and Moses allows for hypothesis recombination between hypotheses that are indistinguishable according to the underlying models.

The key difference between Moses and our system is that, in order to account for target discontinuities, phrases that contains gaps in the target are appended to a partial translation hypothesis in multiple steps. Specifically, each translation hypothesis in our decoder is not only represented as a translation prefix and a coverage set as in Moses, but it also contains a set of isolated phrases (shown in italic in Fig. 3) that must be added to the translation at some later time. For instance, the figure shows how the phrase pair (作出安排, *arrangements ... made*) is being added to a partial translation. The prefix (*arrangements*) is immediately appended to form the hypothesis (*he said arrangements*), and the isolated phrase (*made*) is stored for later use.

Beam search algorithm.

```
1  create initial hypothesis H_∅; add it to S_0^g
2  for j = 0 to J
3    if j > 0 then
4      for n = 1 to N
5        for each H_new in consolidate(H_jn^c)
6          add H_new to S_j^g
7    if j < J then
8      for n = 1 to N
9        H_old := H_jn^g
10       u := first uncovered source word of H_old
11       for m = u to u + distortionLimit
12         for each (s_k, t_k) in translation_options(m)
13           if source s_k does not overlap H_old then
14             H_new := combine(H_old, s_k, t_k)
15             add H_new to S_{j+l}^c, where l = |s_k|
16   return arg max(S_J^g)
```

Table 1: Discontinuous phrase-based MT.

A beam search algorithm for discontinuous phrase-based MT is shown in Table 1. Pruning is done implicitly in the table to avoid cluttering the pseudo-code. The algorithm handles $2J + 1$ stacks $S_0^g, S_1^g, \ldots, S_J^g$ and $S_1^c, \ldots, S_J^c$, where each stack may contain up to N hypotheses $H_{j1}, \ldots, H_{jN}$. The main loop of the algorithm alternates two stages: grow (lines 7–15) and consolidate (lines 3–6).[2] The grow stage is similar to standard phrase-

[1] In order to keep the number of phrases manageable, we additionally require that each (maximal) contiguous substring of s_k and t_k be connected with at least one word alignment.

[2] The distinction between S_i^g and S_i^c stacks ensures that the consolidate operation does not read and write hypotheses on the same stack. While it may seem effective to store hypotheses in

based MT: we take a hypothesis H^g_{jn} from S^g_j and combine it with a translation option (s_k, t_k), which yields a new hypothesis that is added to stack S^c_{j+l} (where $l = |s_k|$). The second stage, consolidate, lets the decoder select any number of isolated phrases (not necessarily all, and possibly zero) and append them in any order at the end of the current translation.[3] Consolidation operations are marked with stars in the figure (for simplicity, the figure does not display consolidations that keep hypotheses unchanged). We limit the number of isolated phrases to 4, which is generally enough to account for most transformations seen in the data. Any hypothesis in the last beam S^g_j is automatically discarded if it contains any isolated phrase.

One last difference with standard decoders is that we also handle source discontinuities. This problem is a known instance of MT by pattern matching (Lopez, 2007), which we already mentioned in the previous section. The function translation_options(m) of Table 1 returns the set of options applicable at position m using this pattern matching algorithm. Since this function is invoked a large number of times, it is important to precompute its return values for each m prior to decoding.

4 Features

Our system incorporates the same eight baseline features of Moses: two relative-frequency phrase translation probabilities $p(e|f)$ and $p(f|e)$, two lexically-weighted phrase translation probabilities (Koehn et al., 2003) $lex(e|f)$ and $lex(f|e)$, a language model probability, word penalty, phrase penalty, and linear distortion, and we optionally add 6 lexicalized reordering features as computed in Moses.

Our computation of linear distortion is different from the one in Moses, since we need to account for discontinuous phrases. We found that it is crucial to penalize discontinuous phrases that have relatively long gaps. Hence, in our computation of linear distortion, we treat continuous subphrases of each discontinuous phrase as if they were continuous phrases on their own. Specifically, let $\bar{\mathbf{s}} = (\bar{s}_1, \ldots, \bar{s}_L)$ be the list of L (maximal) continuous subphrases of the K source phrases ($L \geq K$) selected for a given translation hypothesis. Subphrases in $\bar{\mathbf{s}}$ are enumerated according to their order in the target language, which may be different from the source-side ordering. We then compute the linear distortion between pair of successive elements $(\bar{s}_i, \bar{s}_{i+1})$ as follows:

$$d(\bar{\mathbf{s}}) = \bar{s}_1^{first} + \sum_{i=2}^{L} \left| \bar{s}_{i-1}^{last} + 1 - \bar{s}_i^{first} \right|$$

where the superscripts $first$ and $last$ respectively refer to source position of the first and last word of a given subphrase. Fig. 4 shows an example of how distortion is computed for phrases (s_1, s_2, s_3), including the discontinuous phrase s_2 split into three continuous subphrases. In practice, we compute intra-phrase (shown with thin arrows in the figure) and inter-phrase linear distortion separately in order to produce two distinct features, since translation tends to improves when the intra-phrase cost has a lower feature weight.

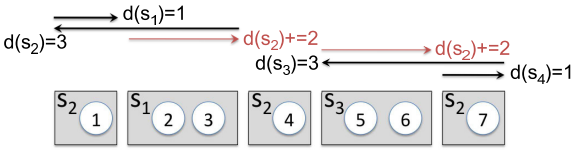


Figure 4: Linear distortion computed using both continuous and discontinuous phrase.

Finally, we add two features that are not present in Moses. First, we penalize target discontinuities by including a feature that is the sum of the lengths of all target gaps. The second feature is the count of discontinuous phrases that are in configurations (cross-serial DTU (Søgaard and Kuhn, 2009) and "bonbon" (Simard et al., 2005)) that can't be handled by 2-SCFG systems. The advantage of such features is two-fold. First, similarly to hierarchical systems, they prevent many distorted reorderings that are unlikely to correspond to quality translations. Second, it imposes soft rather than hard constraints, which means that the decoder is entirely free to violate hierarchical constraints when these violations are supported by other features.

different stacks depending on the number of isolated phrases, we have not found various implementations of this idea to work better than the algorithm described here.

[3] We let isolated phrases be reordered freely, with only three constraints: (1) the internal word order must be preserved, i.e., a phrase may not be split or reordered. (2) isolated phrases drawn from the same discontinuous phrase must appear in the specified order (i.e., the phrase *A ... B ... C* may not yield the translation *A ... C ... B*). (3) Empty gaps are forbidden.

5 Experimental Setup

Three systems are evaluated in this paper: Moses (Koehn et al., 2007), Joshua (Li et al., 2009) – a reimplementation of Hiero, and our phrase-based system. We made our best attempts to make our system comparable to Moses. That is, when no discontinuous phrases are provided to our system, it generates an output that is almost identical to Moses (only about 1% of translations differ on average). In both systems, we use the default settings of Moses, i.e., we set the beam size to 200, the distortion limit to 6, we limit to 20 the number of target phrases that are loaded for each source phrase, and we use the same default eight features of Moses. We use version 1.3 of Joshua with its default settings. Both Moses and our system are evaluated with and without lexicalized reordering (Tillmann, 2004).[4] We believe it to be fair to compare Joshua against phrase-based systems that exploit lexicalized reordering, since Hiero's hierarchical rules are also lexically sensitive.[5]

The language pair for our experiments is Chinese-to-English. The training data consists of about 28 million English words and 23.3 million Chinese words drawn from various news parallel corpora distributed by the Linguistic Data Consortium (LDC). In order to provide experiments comparable to previous work, we used the same corpora as (Wang et al., 2007). We performed word alignment using a cross-EM word aligner (Liang et al., 2006). For this, we ran two iterations of IBM Model 1 and two HMM iterations. Finally, we generated a symmetric word alignment from cross-EM Viterbi alignment using the Moses grow-diag heuristic in the case Moses and our system. In the case of Joshua, we used the grow-diag-final heuristic since this gave better results.

In order to train a competitive baseline given our computational resources, we built a large 5-gram language model using the Xinhua and AFP sections of the Gigaword corpus (LDC2007T40) in addition to the target side of the parallel data. This data represents a total of about 700 million words. We manually removed documents of Gigaword that were released during periods that overlap with those of our development and test sets. The language model was smoothed with the modified Kneser-Ney algorithm as implemented in SRILM (Stolcke, 2002), and we only kept 4-grams and 5-grams that occurred at least three times in the training data.

For tuning and testing, we use the official NIST MT evaluation data for Chinese from 2003 to 2008 (MT03 to MT08), which all have four English references for each input sentence. We used the 1664 sentences of MT06 for tuning and development and all other sets for testing. Parameter tuning was done with minimum error rate training (Och, 2003), which was used to maximize IBM BLEU-4 (Papineni et al., 2001). Since MERT is prone to search errors, especially with large numbers of parameters, we ran each tuning experiment four times with different initial conditions. We used n-best lists of size 200. In the final evaluations, we report results using both TER version 0.7.25 (Snover et al., 2006) and BLEU-4 (both uncased).

6 Results

We start by comparing some translations generated by the best configurations of Joshua, Moses, and our phrase-based decoder, systems we will empirically evaluate later in this section. Fig. 5 shows translations of our development set MT06, which were selected because our system makes a crucial use of discontinuous phrases. In the first example, the Chinese input contains 当 ... 时, which typically translates as *when*. Lacking an entry for the input phrase 当生命权被剥夺时 in its phrase table, Moses is unable to translate this segment appropriately, and must instead split this phrase to generate the translation *when the right was deprived of*, where 时 is translated into *of*. This is evidently a poor translation. Conversely, our system uses a discontinuous phrase to translate 当 ... 时, and translates the intervening words separately.

The remaining three translations all contain cross-serial DTUs (Søgaard and Kuhn, 2009) and thus would be difficult to generate using 2-SCFG systems. The second example motivates the idea

[4] We use Moses' default orientations: monotone, swap, and discontinuous. As far as this reordering model is concerned, we treat discontinuous phrases as continuous, i.e., we simply ignore what lies within gaps to determine phrase orientation.

[5] (Tillmann, 2004) learns for each phrase a tendency to either remain monotone or to swap with other phrases. As noted in (Lopez, 2008), Hiero can represent the same information with hierarchical rules of the form uX, Xu, and XuX. Hiero actually models lexicalized reordering patterns that (Tillmann, 2004) does not account for, e.g., a transformation from X_1uX_2v to $X_2u'v'X_1$.

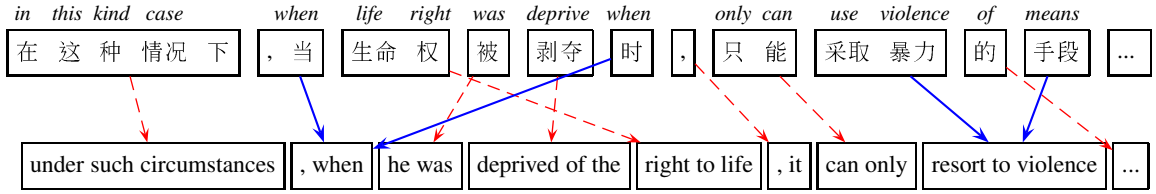
MT06 — segment 1589

Reference: Under such circumstances, when the right of existence was deprived, the only way remaining was to overthrow the existing dynasty by force and try to replace it.

Joshua: Under such circumstances, when life be deprived, can only resort to violence to overthrow the current dynasty, trying to replace,

Moses: Under such circumstances, when the right was deprived of, can only adopt the means of violence, in an attempt to overthrow the present dynasty replaced,

This work: Under such circumstances, when he was deprived of the right to life, it can only resort to violence in an attempt to overthrow the current dynasty replaced,



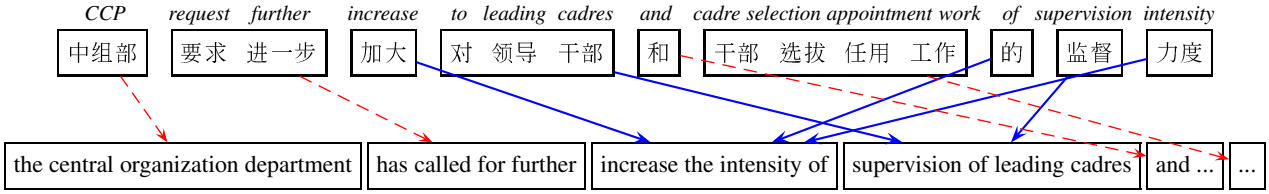
MT06 — segment 1044

Reference: CCP organization ministry demands to further enlarge strength of supervision of leading cadres and cadre selection and appointment

Joshua: Department demands further intensify supervision over the work of selecting and appointing leading cadres, and intensify

Moses: The central organization department, called on leading cadres, further increase the intensity of supervision over work of selecting and appointing cadres.

This work: The central organization department has called for further increase the intensity of supervision of leading cadres and the work of selecting and appointing cadres.



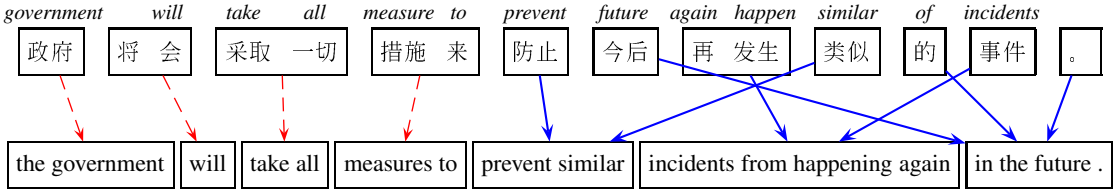
MT06 — segment 559

Reference: The government will take all possible measures to prevent similar incidents from happening in the future.

Joshua: Government will take all measures to prevent the re-occurrence of similar incidents in the future.

Moses: The government will take all measures to prevent the occurrence of similar incidents in the future.

This work: The government will take all measures to prevent similar incidents from happening again in the future.



MT06 — segment 769

Reference: He also said that the arrangements are being made now for the visits.

Joshua: He also said that now is making arrangements for this visit.

Moses: He also said that the current visit is to make arrangements.

This work: He also said that the current arrangements are made for the visit.

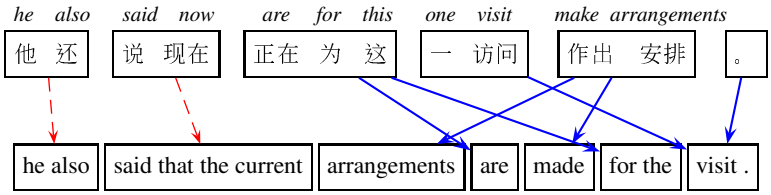


Figure 5: Actual translations produced by Joshua, Moses, and our system. For our system, we also display phrase alignments, including discontinuous phrase alignments. Results for these three systems here are displayed in rows 2, 4, and 8 of Table 2. The thick blue arrows represent alignments between discontinuous phrases, while red segmented arrows align continuous phrases.

	System	Gaps	LexR	MT06 (tune) BLEU	MT06 (tune) TER	MT03 BLEU	MT03 TER	MT04 BLEU	MT04 TER	MT05 BLEU	MT05 TER	MT08 BLEU	MT08 TER	ALL BLEU	ALL TER
1	hierarchical	src	yes	33.55	58.04	33.25	59.73	36.03	58.92	32.03	61.11	26.30	61.30	31.70	58.21
2	(Joshua)	src+tgt	yes	33.84	58.11	33.47	59.85	36.10	58.82	32.17	61.20	26.61	61.21	31.90	58.22
3	phrase-based	no	no	33.17	59.24	32.60	60.80	35.38	59.55	31.15	62.43	25.56	61.98	31.08	59.14
4	(Moses)	no	yes	34.25	58.23	33.72	60.42	36.37	59.18	32.49	61.80	26.70	61.48	32.16	58.56
5	discontinuous phrase-based (this work)	src	no	33.77	58.56	33.20	60.42	36.17	59.13	31.75	61.62	25.99	61.47	31.68	58.60
6		tgt	no	33.27	58.98	32.95	60.42	35.41	59.35	31.08	62.45	25.69	61.71	31.17	58.93
7		src+tgt	no	33.86	58.26	33.32	60.02	36.36	58.56	31.87	61.35	26.13	61.29	31.81	58.25
8		src+tgt	yes	35.00	56.85	34.96	57.97	37.44	57.61	33.39	59.92	26.74	60.51	32.93	57.03
Improvement over hierarchical				**+1.16**	**−1.26**	**+1.49**	**−1.88**	**+1.34**	**−1.21**	**+1.22**	**−1.28**	+0.13	**−0.70**	**+1.03**	**−1.19**
Improvement over phrase-based				**+0.75**	**−1.38**	**+1.24**	**−2.45**	**+1.07**	**−1.57**	**+0.90**	**−1.88**	+0.04	**−0.97**	**+0.77**	**−1.53**
Number of sentences				*1664*		*919*		*1788*		*1082*		*1357*		*6810*	

Table 2: Our system compared again conventional and hierarchical phrase-based MT (Moses and Joshua). using uncased BLEUr4n4[%] and TER[%]. LexR indicates whether lexicalized reordering is enabled or not. We use randomization tests (Riezler and Maxwell, 2005) to determine significance of our best results (row 8) against Joshua (row 2) and Moses (row 4): differences marked in bold are significant at the $p \leq .01$ level.

that larger translation units, including discontinuous phrases, lead to better translations. The reference includes the translation *enlarge strength of supervision of leading cadres*, and our system is able to produce a translation that is almost identical (*increase the intensity of supervision of leading cadres*) using only two phrases, pulling together input words that are fairly far apart in the sentence. The third Chinese sentence has a word order quite different from English, but our decoder flexibly reorders it in a manner that can't be handled with SCFG decoders to give a word order (*prevent similar events from happening*) that matches the one in the reference. The last Chinese sentence includes the topicalization word 为 (*for*), which indicates the input sentence has no subject. One way to properly handle this translation is to turn the sentence into a passive in English (as in the reference), a transformation our system does, thanks to its support for complex reorderings.

Our main results are displayed in Table 2. First, Joshua systematically outperforms the Moses baseline (+0.82 BLEU point and −0.92 TER point on average), but performance of the two is about the same when Moses incorporates lexicalized reordering. This finding is consistent with previous work (Lopez, 2008). The results of our system displayed in rows 5–8 demonstrate that our system consistently outperforms Moses, whether they both use lexicalized reordering or not. The performance of our best system—i.e., with lexicalized reordering and both source and target gaps—is significantly better than the best Moses system (+0.77 BLEU and −1.53 TER). While the performance of our system without lexicalized reordering is close to that of Joshua, our system with lexicalized reordering significantly outperforms Joshua ($p \leq .01$) in 9 out of 10 evaluations. The single experiment where our improvement over Hiero is insignificant (i.e., BLEU on MT08) is mainly affected by a discrepancy of length (our brevity penalty on MT08 is 0.92).

It is interesting to notice that our system allowing phrasal discontinuities only on the source (row 5) performs almost as well as the system that allows them on both sides (row 7). For instance, while source discontinuities improve performance by 0.7 BLEU point on MT06, further enabling target discontinuities only raises performance by a mere 0.09 BLEU point. This naturally raises the question of whether our support for target gaps is ineffective, or whether target-discontinuous phrases are somewhat superfluous to the MT task. While it is certainly difficult to either confirm or deny the latter hypothesis, we can at least compare our handling of target-discontinuous phrases with hierarchical systems. In one additional set of experiments, we removed target-discontinuous phrases in Joshua prior to MERT and test time. Specifically, we removed all hierarchical phrases whose target side has the form uXv, $uXvX$, $XuXv$, and $uXvXw$, and only allowed rules whose target side has the form uX, Xu, XuX, XXu, or uXX. After this filtering, we found that target-discontinuous phrases in Joshua are also not crucial to its performance, since their removal only caused a drop of 0.2 BLEU point (row 1) and almost no change in terms of TER. We speculate that using target discontinuous phrases is more diffi-

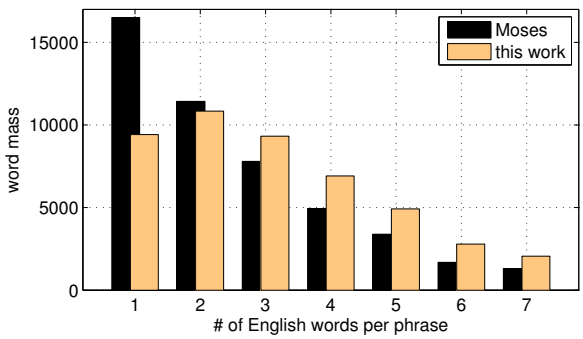


Figure 6: Phrase length histogram for MT06.

cult, since it represents a generation rather than just a matching problem.

In this paper, we have also argued that a main benefit of discontinuous phrases—and particularly source-discontinuous phrases—is that the decoder is allowed to use larger translation units than when restricted to continuous phrases. This claim is confirmed in Fig. 6. We find that our decoder makes effective use of the extended set of translation options at its disposal: While the Moses baseline translates MT06 with an average 1.73 words per phrase, adding support for discontinuities increases this average to 2.16, and reduces by 43% the use of single word phrases. On MT06, 53% of the translated sentences produced by our best system use at least one source-discontinuous phrase, and 9% of them exploit one or more target-discontinuous phrases.

7 Related Work

The main goal of this paper is to show that discontinuous phrases can greatly improve the performance of phrase-based systems. While some of the most recent phrase-based systems (Chiang, 2007; Watanabe et al., 2006) exploit context-free decoding algorithms (CKY, Earley, etc.) to cope with discontinuities, our system preserves the simplicity and speed of conventional phrase-based decoders, and in particular does not build any intermediate tree structure, does not impose any hard reordering constraints other than the distortion limit, and still achieves translation performance that is superior to that of a state-of-the-art hierarchical system.

A few previous non-hierarchical systems have also exploited phrasal discontinuities. The most notable previous attempt to incorporate gaps is described in (Simard et al., 2005). Simard et al. presents an extension to Moses that allows gaps in both source and target phrases, though each of their gap symbols must span exactly one word. This fact makes decoding simpler, since the position of all target words in a translation hypothesis is known as soon as the hypothesis is laid down, but fixed-size discontinuous phrases are less general and increase sparsity. By comparison, our gaps may span any number of words, so we have an increased ability to flexibly match the input sentence effectively. (Crego and Yvon, 2009) also handles gaps, though this work is applicable to an n-gram-based SMT framework (Mariòo et al., 2006), which is fairly different from the phrase-based framework.

8 Conclusions

In this paper, we presented a generalization of conventional phrase-based decoding to handle discontinuities in both source and target phrases. Our system significantly outperforms Moses and Joshua, two standard implementations of conventional and hierarchical phrase-based decoding. We found that allowing discontinuities in the source is more useful than target discontinuities in our system, though we found that this turns out to also be the case with the hierarchical phrases of Joshua. In future work, we plan to extend the parameterization of phrase-based lexicalized reordering models to be sensitive to these discontinuities, and we will also consider adding syntactic features to our models to penalize discontinuities that are not syntactically motivated (Marton and Resnik, 2008; Chiang et al., 2009). The discontinuous phrase-based MT system described in this work is part of Phrasal, an open-source phrase-based system available for download at http://nlp.stanford.edu/software/phrasal.

Acknowledgements

The authors thank three anonymous reviewers, Dan Jurafsky, Spence Green, Steven Bethard, Daniel Cer, Chris Callison-Burch, and Pi-Chuan Chang for their helpful comments. This paper is based on work funded by the Defense Advanced Research Projects Agency through IBM. The content does not necessarily reflect the views of the U.S. Government, and no official endorsement should be inferred.

References

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. Scaling phrase-based statistical machine translation to larger corpora and longer phrases. In *Proc. of ACL*, pages 255–262.

Daniel Cer, Michel Galley, Dan Jurafsky, and Christopher Manning. 2010. Phrasal: A statistical machine translation toolkit for exploring new model features. In *Proc. of NAACL-HLT, Demonstration Session*.

David Chiang, Kevin Knight, and Wei Wang. 2009. 11,001 new features for statistical machine translation. In *Proc. of NAACL*, pages 218–226.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Josep Crego and François Yvon. 2009. Gappy translation units under left-to-right SMT decoding. In *Proc. of EAMT*.

Liang Huang, Hao Zhang, and Daniel Gildea. 2005. Machine translation as lexicalized parsing with hooks. In *Proc. of the Ninth International Workshop on Parsing Technology*, pages 65–73.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL*, pages 48–54.

Philipp Koehn, Hieu Hoang, Alexandra Birch Mayne, Christopher 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 *Proc. of ACL, Demonstration Session*.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proc. of AMTA*, pages 115–124.

Zhifei Li, Chris Callison-Burch, Chris Dyer, Juri Ganitkevitch, Sanjeev Khudanpur, Lane Schwartz, Wren N. G. Thornton, Jonathan Weese, and Omar F. Zaidan. 2009. Joshua: an open source toolkit for parsing-based MT. In *Proc. of WMT*.

Percy Liang, Ben Taskar, and Dan Klein. 2006. Alignment by agreement. In *Proc. of HLT-NAACL*, pages 104–111.

Adam Lopez. 2007. Hierarchical phrase-based translation with suffix arrays. In *Proc. of EMNLP-CoNLL*, pages 976–985.

Adam Lopez. 2008. Tera-scale translation models via pattern matching. In *Proc. of COLING*.

José B. Mariòo, Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. N-gram-based machine translation. *Computational Linguistics*, 32(4):527–549.

Yuval Marton and Philip Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *Proc. of ACL*, pages 1003–1011.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Franz Josef Och. 2003. Minimum error rate training for statistical machine translation. In *Proc. of ACL*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*.

Stefan Riezler and John T. Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proc. of Workshop on Evaluation Measures*, pages 57–64.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with non-contiguous phrases. In *Proc. of HLT-EMNLP*, pages 755–762.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA*, pages 223–231.

Anders Søgaard and Jonas Kuhn. 2009. Empirical lower bounds on alignment error rates in syntax-based machine translation. In *Proc. of the Third Workshop on Syntax and Structure in Statistical Translation (SSST-3) at NAACL HLT 2009*, pages 19–27.

Anders Søgaard and Dekai Wu. 2009. Empirical lower bounds on translation unit error rate for the full class of inversion transduction grammars. In *Proc. of IWPT*, pages 33–36.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, pages 901–904.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proc. of HLT-NAACL*, pages 101–104.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proc. of EMNLP-CoNLL*.

Taro Watanabe, Hajime Tsukada, and Hideki Isozaki. 2006. Left-to-right target generation for hierarchical phrase-based translation. In *Proc. of ACL*.

Benjamin Wellington, Sonjia Waxmonsky, and I. Dan Melamed. 2006. Empirical lower bounds on the complexity of translational equivalence. In *Proc. of COLING-ACL*, pages 977–984.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–404.

Ying Zhang and Stephan Vogel. 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proc. of EAMT*.