Improved Models of Distortion Cost for Statistical Machine Translation

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

Abstract

The distortion cost function used in Moses-style machine translation systems has two flaws. First, it does not estimate the future cost of known required moves, thus increasing search errors. Second, all distortion is penalized linearly, even when appropriate reorderings are performed. Because the cost function does not effectively constrain search, translation quality decreases at higher distortion limits, which are often needed when translating between languages of different typologies such as Arabic and English. To address these problems, we introduce a method for estimating future linear distortion cost, and a new discriminative distortion model that predicts word movement during translation. In combination, these extensions give a statistically significant improvement over a baseline distortion parameterization. When we triple the distortion limit, our model achieves a +2.32 BLEU average gain over Moses.

1 Introduction

It is well-known that translation performance in Moses-style (Koehn et al., 2007) machine translation (MT) systems deteriorates when high distortion is allowed. The *linear distortion cost model* used in these systems is partly at fault. It includes no estimate of future distortion cost, thereby increasing the risk of search errors. Linear distortion also penalizes all re-orderings equally, even when appropriate re-orderings are performed. Because linear distortion, which is a soft constraint, does not effectively constrain search, a *distortion limit* is imposed on the translation model. But hard constraints are ultimately undesirable since they prune the search space. For languages with very different word orders in which significant re-ordering is required, the distortion limit can eliminate the *oracle*, or "best," translation prior to search, placing an artificial limit on translation performance (Auli et al., 2009).

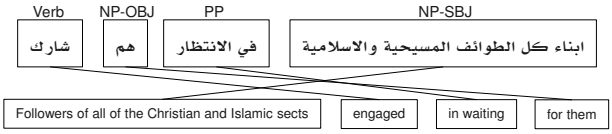


Figure 1: The oracle translation for this Arabic VOS sentence would be pruned during search using typical distortion parameters. The Arabic phrases read right-to-left, but we have ordered the sentence from left-to-right in order to clearly illustrate the re-ordering problem.

To illustrate this problem, consider the Arabic-English example in Figure 1. Assuming that the English translation is constructed left-to-right, the verb شارك *shaaraka* must be translated after the noun phrase (NP) subject. If P phrases are used to translate the Arabic source s to the English target t, then the (unsigned) linear distortion is given by

$$D(s,t) = p^1_{first} + \sum_{i=2}^{P} \left| p^{i-1}_{last} + 1 - p^i_{first} \right| \quad (1)$$

where p_{first} and p_{last} are the first and last source word indices, respectively, in phrase i. By this formula, the cost of the step to translate the NP subject before the verb is 9, which is high relative to the monotone translation path. Moreover, a conventional distortion limit (e.g., 5) would likely force translation of the verb prior to the full subject unless the exact subject phrase existed in the phrase table.[1] Therefore, the correct re-ordering is either improbable or impossible, depending on the choice of distortion parameters.

[1] Our constrained NIST MT09 Arabic-English system, which placed second, used a limit of 5 (Galley et al., 2009).

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 867–875,
Los Angeles, California, June 2010.

The objective of this work is to develop a distortion cost model that allows the distortion limit to be raised significantly without a catastrophic decrease in performance. We first describe an admissible future cost heuristic for linear distortion that restores baseline performance at high distortion limits. Then we add a feature-rich discriminative distortion model that captures e.g. the tendency of Arabic verbs to move right during translation to English. Model parameters are learned from automatic bitext alignments. Together these two extensions allow us to triple the distortion limit in our NIST MT09 Arabic-English system while maintaining a statistically significant improvement over the low distortion baseline. At the high distortion limit, we also show a +2.32 BLEU average gain over Moses with an equivalent distortion parameterization.

2 Background

2.1 Search in Phrase-based MT

Given a J token source input string $\boldsymbol{f} = \left\{f_i^J\right\}$, we seek the most probable I token translation $\boldsymbol{e} = \left\{e_i^I\right\}$. The Moses phrase-based decoder models the posterior probability $p_\lambda\left(e_1^I|f_1^J\right)$ directly according to a log-linear model (Och and Ney, 2004), which gives the decision rule

$$\hat{e} = \arg\max_{I,e_1^I} \left\{ \sum_{m=1}^{M} \lambda_m h_m\left(e_1^I, f_1^J\right) \right\}$$

where $h_m\left(e_1^I, f_1^J\right)$ are M arbitrary feature functions over sentence pairs, and λ_m are feature weights set using a discriminative training method like MERT (Och, 2003). This search is made tractable by the use of beams (Koehn et al., 2003). Hypotheses are pruned from the beams according the sum of the current model score and a future cost estimate for the uncovered source words. Since the number of reordering possibilities for those words is very large—in theory it is exponential—an inadmissible heuristic is typically used to estimate future cost. The baseline distortion cost model is a weighted feature in this framework and affects beam pruning only through the current model score.

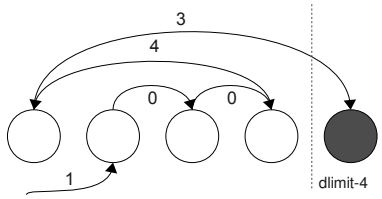
When we say *linear distortion*, we refer to the "simple distortion model" of Koehn et al. (2003) that is shown in Equation (1) and is converted to a cost by multiplying by -1. When extended to phrases, the key property of this model is that monotone decoding gives the least costly translation path. Reorderings internal to extracted phrases are not penalized. In practice, we commonly see n-best lists of hypotheses with linear distortion costs equal to zero. More sophisticated local phrase re-ordering models have been proposed (Tillmann, 2004; Zens and Ney, 2006; Koehn et al., 2007; Galley and Manning, 2008), but these are typically used in addition to linear distortion.

2.2 Arabic Linguistic Essentials

In this paper we use Arabic-English as a case study since we possess a strong experimental baseline. But we expect that the technique presented could be even more effective for high distortion language pairs such as Chinese-English and Hindi-English. Since the analysis that follows is framed in terms of Arabic, we point out several linguistic features that motivate our approach. From the perspective of the three criteria used to specify basic word order typology (Greenberg, 1966), Arabic is somewhat unusual in its combination of features: it has prepositions (not postpositions), adjectives post-modify nouns, and the basic word order is VSO, but SVO and VOS configurations also appear.

The implications for translation to English are: (1) prepositions remain in place, (2) NPs are inverted, and most importantly, (3) basic syntactic constituents must often be identified and precisely re-ordered. The VOS configuration is especially challenging for Arabic-English MT. It usually appears when the direct object is short—e.g., pronominal—and the subject is long. For example, translation of the VOS sentence in Figure 1 requires both a high distortion limit to accommodate the subject movement and tight restrictions on the movement of the PP. The particularity of these requirements in Arabic and other languages, and the difficulty of modeling them in phrase-based systems, has inspired significant work in source language preprocessing (Collins et al., 2005; Habash and Sadat, 2006; Habash, 2007).

Finally, we observe that target language models cannot always select appropriate translations when basic word order transformation is required. By not modeling source side features like agreement—which, in Arabic, appears between both verb and



step k	F_k	Δ_{cost}	$D(s,t)$	$D(s,t)+\Delta_{cost}$
0	3	3	1	4
1	5	2	0	2
2	7	2	0	2
3	0	−7	4	−3
4	0	0	3	3
			8	**8**

Figure 2: Translation sequence in which the distortion limit is reached and the decoder is forced to cover the first skipped word. Future cost estimation penalizes the two monotone steps, yet total distortion cost remains unchanged.

subject, and adjective and noun—baseline phrase-based systems rely on the language model to specify an appropriate target word order (Avramidis and Koehn, 2008). Returning to Figure 1, we could have an alternate hypothesis *They waited for the followers of the Christian and Islamic sects*, which is acceptable English and has low distortion, but is semantically inconsistent with the Arabic.

	dlimit = 5	dlimit = 15
LINEAR	51.65	**49.35**
DEGREE 1.5	**51.69 (+0.04)**	48.73 (−0.62)
DEGREE 2	51.55 (−0.10)	48.40 (−0.95)

Table 1: BLEU-4 [%] dev set (MT04) scores (uncased) for several polynomial distortion models. Higher degree polynomial distortion models underperform at a high distortion limit (15).

3 The Cost Model

In this section we describe the new distortion cost model, which has four independent components.

3.1 Future Cost Estimation

Despite its lack of sophistication, linear distortion is a surprisingly effective baseline cost model for phrase-based MT systems. It can be computed in constant time, gives non-decreasing values that are good for search, and does not require an ancillary feature to adjust for the number of components in the calculation (e.g., language model scores are adjusted by the word penalty). Moreover, when a large training bitext is used, many local re-orderings are captured in the phrase table, so the decoder can often realize competitive performance by finding a best set of phrases with low distortion. But linear distortion is not the only unlexicalized alternative: we can use any function of the jump width. Table 1 shows development set (MT04) performance for polynomials of degree 1.5 and degree 2. The linear model is more effective than the higher order functions, especially at a higher distortion limit.

Nevertheless, Table 1 shows an unacceptable decrease in translation performance at the high distortion limit for all three polynomial models. In Moses, the reason is due in part to a dramatic underestimation of future re-ordering cost. Consider Figure 2 in which a distortion limit of 4 is used. The first word is skipped, and translation proceeds monotonically until the distortion limit forces the decoder to cover the first word. At low distortion limits, single phrases often saturate the distortion window, so underestimation is not problematic. But at high distortion limits, the decoder can skip many source positions at low cost before the search is constrained by the distortion limit. Words and phrases sprinkled carelessly throughout the hypotheses are evidence of errant search directions that have not been appropriately penalized by the distortion cost model.

To constrain search, we add an admissible future cost estimate to the linear model.[2] By definition, the model has a least cost translation path: monotone. Therefore, we can add to the baseline calculation $D(s,t)$ the cost of skipping back to the first uncovered source word and then translating the remaining positions monotonically. It can be verified by induction on $|\boldsymbol{C}|$ that this is an admissible heuristic.

Formally, let j represent the first uncovered index in the source coverage set $\boldsymbol{C}$. Let $\boldsymbol{C}_j$ represent the subset of $\boldsymbol{C}$ starting from position j. Finally, let j' represent the leftmost position in phrase p applied at translation step k. Then the future cost estimate F_k

[2]Moore and Quirk (2007) propose an alternate future cost formulation. However, their model seems prone to the same deterioration in performance shown in Table 1. They observed decreased translation quality above a distortion limit of 5.

is

$$F_k = \begin{cases} |\boldsymbol{C}_j| + (j' + |p| + 1 - j) & \text{if } j' > j \\ 0 & \text{otherwise} \end{cases}$$

For $k > 0$, we add the difference between the current future cost estimate and the previous cost estimate $\Delta_{cost} = F_k - F_{k-1}$ to the linear penalty $D(s,t)$.[3] Table 2 shows that, as expected, the difference between the baseline and augmented models is statistically insignificant at a low distortion limit. However, at a very high distortion limit, the future cost estimate approximately restores baseline performance. While we still need a distortion limit for computational efficiency, it is no longer required to improve translation quality.

3.2 A Discriminative Distortion Model

So far, we have developed a search heuristic function that gives us a greater ability to control search at high distortion limits. Now we need a cost model that is sensitive to the behavior of certain words during translation. The model must accommodate a potentially large number of overlapping source-side features defined over the (possibly whole) translation sequence. Since we intend to train on automatic word alignments, data sparsity and noise are also risks. These requirements motivate two choices. First, we use a discriminative log-linear framework that predicts one of the nine discretized distortion classes in Figure 3. Let $d_{j,j'}$ indicate the class corresponding to a jump from source word j to j' computed as $(j + 1 - j')$. The discriminative distortion classifier is then

$$p_\lambda\left(d_{j,j'} | f_1^J, j, j'\right) = \frac{\exp\left[\sum_{m=1}^M \lambda_m h_m\left(f_1^J, j, j', d_{j,j'}\right)\right]}{\sum_{d^i_{j,j'}} \exp\left[\sum_{m=1}^M \lambda_m h_m\left(f_1^J, j, j', d^i_{j,j'}\right)\right]}$$

where λ_m are feature weights for the $h_m(f_1^J, j, j', d^i_{j,j'})$ arbitrary feature functions. This log conditional objective function is convex and can be optimized with e.g. a gradient-based procedure.

[3]One implementation choice is to estimate future cost to an artificial end-of-sentence token. Here the decoder incurs a penalty for covering the last word prior to completing a hypothesis. Although this implementation is inconsistent with Moses linear distortion, we find that it gives a small improvement.

	dlimit = 5	dlimit = 15
BASELINE	51.65	49.35
FUTURECOST	51.73	51.65

Table 2: BLEU-4 [%] dev set scores (uncased) for the linear distortion with future cost estimation.

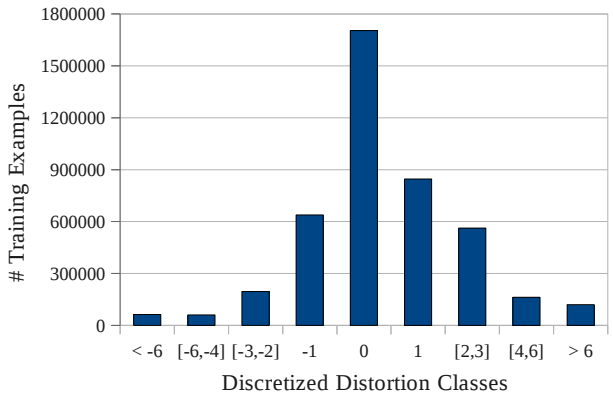


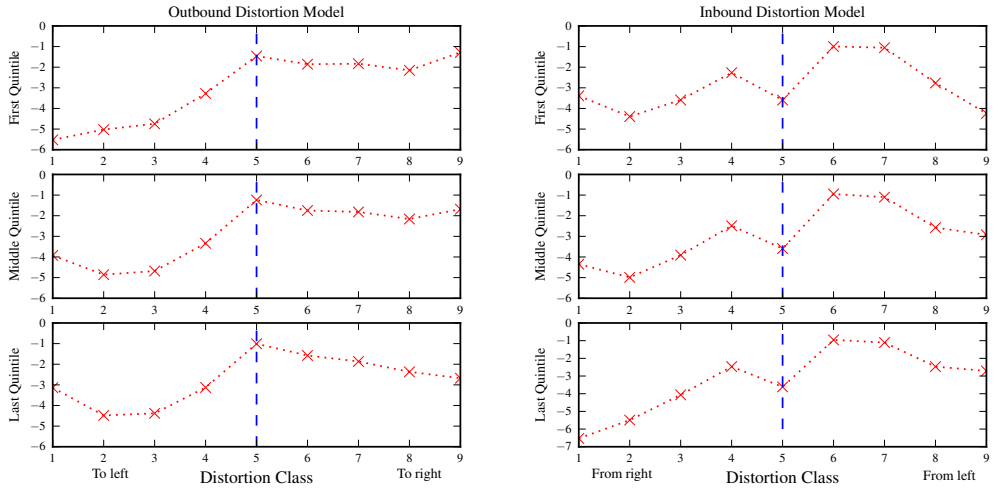
Figure 3: Distortion in Arabic-English translation is largely monotonic, but with noticeable right movement as verbs move around arguments and nouns around modifiers. The ability to predict movement decreases with the jump size, hence the increasing bin boundaries.

Second, we expect that many words will not be useful for predicting translation order.[4] In a large training bitext, it can be extremely tedious to identify informative words and word classes analytically. Our final decision is then to optimize the parameter weights λ_m using L_1 regularization (Andrew and Gao, 2007), a technique that can learn good models in the presence of many irrelevant features.[5] The L_1 regularizer saves us from filtering the training data (e.g., by discarding all words that appear less than an empirically-specified threshold), and provides sparse feature vectors that can be analyzed separately during feature engineering.

We train two independent distortion models. For a transition from source word j to j', we learn an *outbound* model in which features are defined with respect to word j. We have a corresponding *inbound*

[4]To train the models, we inverted and sorted the intersection alignments in the bitext. In our baseline system, we observed no decrease in performance between intersection and e.g. grow-diag. However we do expect that our method could be extended to multi-word alignments.

[5]We also add a Gaussian prior $p(\lambda) \backsim \mathcal{N}(0,1)$ to the objective (Chen and Rosenfeld, 1999). Using both L_1 and L_2 regularization is mathematically odd, but often helps in practice.



(a) شارك / VBD *shaaraka* (“he engaged”) (b) الامريكي / JJ *al-aamriikii* (“American”)

Figure 4: Selected discriminative cost curves (log scale) over three quintiles of the relative position feature. We condition on the word, POS, and length features. The classes correspond to those shown in Figure 3. (4a) The VSO basic word order is evident: early in the sentence, there is a strong tendency towards right movement around arguments after covering the verb. However, right movement is increasingly penalized at the end of the sentence. (4b) Adjectives post-modify nouns, so the model learns high inbound probabilities for jumps from positions earlier in the sentence. However, the curve is bi-modal reflecting right inbound moves from other adjectives in NPs with multiple modifiers.

model trained on features with respect to j'. At training time, we also add sentence beginning and ending delimiters such that inbound probabilities are learned for words that begin sentences (e.g., nouns) and outbound probabilities are available for tokens that end sentences (e.g., punctuation).

As a baseline, we use the following binary features: words, part-of-speech (POS) tags, relative source sentence position, and source sentence length. Relative source sentence position is discretized into five bins, one for each quintile of the sentence. Source sentence length is divided into four bins with bounds set empirically such that training examples are distributed evenly. To simplify the decoder integration for this evaluation, we have chosen context-free features, but the framework permits many other promising possibilities such as agreement morphology and POS tag chains.

Our models reveal principled cost curves for specific words (Figure 4). However, monotonic decoding no longer gives the least costly translation path, thus complicating future cost estimation. We would need to evaluate all possible re-orderings within the k-word distortion window. For an input sentence of length n, Zens (2008) shows that the number of reordering possibilities r_n is

$$r_n = \begin{cases} k^{n-k} \cdot k! & n > k \\ n! & n \leq k \end{cases}$$

which has an asymptotic complexity $\Theta(k^n)$. Instead of using an inadmissible heuristic as is done in beam pruning, we take a shortcut: we include the linear future cost model as a separate feature. Then we add the two discriminative distortion features, which calculate the inbound and outbound log probabilities of the word alignments in a hypothesis. Since hypotheses may have different numbers of alignments, we also include an alignment penalty that adjusts the discriminative distortion scores for unaligned source words. The implementation and behavior of the alignment penalty is analogous to that of the word penalty. In total, the new distortion cost model has four independent MT features.

4 MT Evaluation

4.1 Experimental Setup

Our MT system is Phrasal (Cer et al., 2010), which is a Java re-implementation of the Moses

dlimit = 5	MT03	MT05	MT06	MT08	Avg
MOSESLINEAR	52.31	52.67	42.97	41.29	
COUNTS	52.05	52.32	42.28	40.56	
FUTURE	52.26 (−0.05)	52.53 (−0.14)	43.04 (+0.07)	41.01 (−0.28)	−0.09
DISCRIM+FUTURE	**52.68* (+0.37)**	**53.13* (+0.46)**	**43.75** (+0.78)**	**41.82** (+0.53)**	+0.59

Table 3: BLEU-4 [%] scores (uncased) at the distortion limit (5) used in our baseline NIST MT09 Arabic-English system (Galley et al., 2009). **Avg** is a weighted average of the performance deltas. The stars for positive results indicate statistical significance compared to the MOSESLINEAR baseline (*: significance at $p \leq 0.05$; **: significance at $p \leq 0.01$)

dlimit = 15	MT03	MT05	MT06	MT08	Avg
MOSESLINEAR	51.04	51.35	41.01	38.83	
COUNTS	49.92	49.73	39.44	37.65	
LEX	50.96	51.21	41.87	39.38	
FUTURE	52.28** (+1.24)	52.45** (+1.10)	42.78** (+1.77)	41.01** (+2.18)	+1.66
DISCRIM+FUTURE	**52.36** (+1.32)**	**53.05** (+1.70)**	**43.65** (+2.64)**	**41.68** (+2.85)**	+2.32
num. sentences	*663*	*1056*	*1797*	*1360*	*4876*

Table 4: BLEU-4 [%] scores (uncased) at a very high distortion limit (15). DISCRIM+FUTURE also achieves a statistically significant gain over the MOSESLINEAR dlimit=5 baseline for MT05 ($p \leq 0.06$), MT06 ($p \leq 0.01$), and MT08 ($p \leq 0.01$).

decoder with the same standard features: four translation features (phrase-based translation probabilities and lexically-weighted probabilities), word penalty, phrase penalty, linear distortion, and language model score. We disable baseline linear distortion when evaluating the other distortion cost models. To tune parameters, we run MERT with the Downhill Simplex algorithm on the MT04 dataset. For all models, we use 20 random starting points and generate 300-best lists.

We use the NIST MT09 constrained track training data, but remove the UN and comparable data.[6] The reduced training bitext has 181k aligned sentences with 6.20M English and 5.73M Arabic tokens. We create word alignments using the Berkeley Aligner (Liang et al., 2006) and take the intersection of the alignments in both directions. Phrase pairs with a maximum target or source length of 7 tokens are extracted using the method of Och and Ney (2004).

We build a 5-gram language model from the Xinhua and AFP sections of the Gigaword corpus (LDC2007T40), in addition to all of the target side training data permissible in the NIST MT09 constrained competition. We manually remove Gigaword documents that were released during periods that overlapped with the development and test sets. The language model is smoothed with the modified Kneser-Ney algorithm, retaining only trigrams, 4-grams, and 5-grams that occurred two, three, and three times, respectively, in the training data.

We remove from the test sets source tokens not present in the phrase tables. For the discriminative distortion models, we tag the pre-processed input using the log-linear POS tagger of Toutanova et al. (2003). After decoding, we strip any punctuation that appears at the beginning of a translation.

4.2 Results

In Table 3 we report uncased BLEU-4 (Papineni et al., 2001) scores at the distortion limit (5) of our most competitive baseline Arabic-English system. MOSESLINEAR uses the linear distortion model present in Moses. COUNTS is a separate baseline with a discrete cost model that uses unlexicalized maximum likelihood estimates for the same classes present in the discriminative model. To show the effect of the components in our combined distortion model, we give separate results for linear distortion with future cost estimation (FUTURE) and for the combined discriminative distortion model (DISCRIM+FUTURE) with all four features: linear distortion with future cost, inbound and outbound proba-

[6] Removal of the UN data does not affect the baseline at a distortion limit of 5, and lowers the higher distortion baseline by −1.40 BLEU. The NIST MT09 data is available at http://www.itl.nist.gov/iad/mig/tests/mt/2009/.

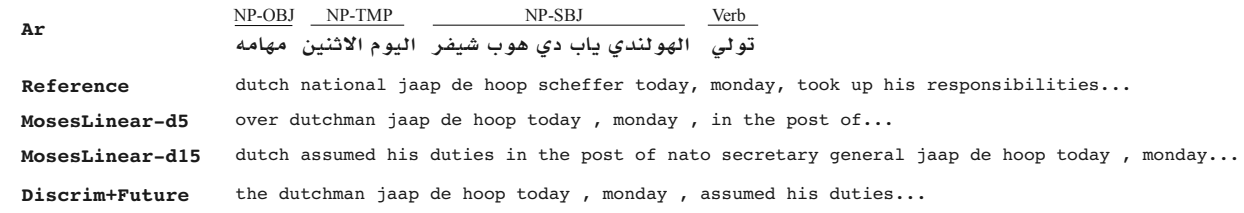


Figure 5: Verb movement around both the subject and temporal NPs is impossible at a distortion limit of 5 (MOSESLINEAR-d5). The baseline system at a high distortion limit mangles the translation (MOSESLINEAR-d15). DISCRIM+FUTURE (dlimit=15) correctly guides the search. The Arabic source is written right-to-left.

bilities, and the alignment penalty.

The main objective of this paper is to improve performance at very high distortion limits. Table 4 shows performance at a distortion limit of 15. To the set of baselines we add LEX, which is the lexicalized re-ordering model of Galley and Manning (2008). This model was shown to outperform other lexicalized re-ordering models in common use.

Statistical significance was computed with the approximate randomization test of Riezler and Maxwell (2005), which is less sensitive to Type I errors than bootstrap re-sampling (Koehn, 2004).

5 Discussion

The new distortion cost model allows us to triple the distortion limit while maintaining a statistically significant improvement over the MOSESLINEAR baseline at the lower distortion limit for three of the four test sets. More importantly, we can raise the distortion limit in the DISCRIM+FUTURE configuration at minimal cost: a statistically insignificant -0.2 BLEU performance decrease on average. We also see a considerable improvement over both the MOSESLINEAR and LEX baselines at the high distortion limit (Figure 5). As expected, future cost estimation alone does not increase performance at the lower distortion limit.

We also observe that the effect of conditioning on evidence is significant: the COUNTS model is categorically worse than all other models. To understand why, we randomly sampled 500 sentences from the excluded UN data and computed the log-likelihoods of the alignments according to the different models.[7] In this test, COUNTS is clearly better with a score of -23388 versus, for example, the inbound model at -38244. The explanation is due in part to optimization. The two discriminative models often give very low probabilities for the outermost classes. Noise in the alignments along with the few cases of long-distance movement are penalized heavily. For Arabic, this property works in our favor as we do not want extreme movement (as we might with Chinese or German). But COUNTS applies a uniform penalty for all movement that exceeds the outermost class boundaries, making it more prone to search errors than even linear distortion despite its favorable performance when tested in isolation.

Finally, we note that previous attempts to improve re-ordering during search (particularly long-distance re-ordering (Chiang, 2007)) have delivered remarkable gains for languages like Chinese, but improvements for Arabic have been less exceptional. By relaxing the distortion limit, we have left room for more sophisticated re-ordering models in conventional phrase-based decoders while maintaining a significant performance advantage over hierarchical systems (Marton and Resnik, 2008).

6 Prior Work

There is an expansive literature on re-ordering in statistical MT. We first review the development of re-ordering constraints, then describe previous cost models for those constraints in beam search decoders. Because we allow re-ordering during search, we omit discussion of the many different methods for preprocessing the source input prior to monotonic translation. Likewise, we do not recite prior work in re-ranking translations.

Re-ordering constraints were first introduced by Berger et al. (1996) in the context of the IBM translation models. The *IBM constraints* treat the source

[7] We approximated linear distortion using a Laplacian distribution with estimated parameters $\hat{\mu} = 0.51$ and $\hat{b} = 1.76$ (Goodman, 2004).

word sequence as a coverage set $\boldsymbol{C}$ that is processed sequentially. A source token is "covered" when it is aligned with a new target token. For a fixed value of k, we may leave up to $k-1$ positions uncovered and return to them later. We can alter the constraint slightly such that for the first uncovered position $u \notin \boldsymbol{C}$ we can cover position j when

$$j - u < k \qquad j \notin \boldsymbol{C}$$

which is the definition of the distortion limit used in Moses. Variations of the IBM constraints also exist (Kanthak et al., 2005), as do entirely different regimes like the hierarchical *ITG constraints*, which represent the source as a sequence of blocks that can be iteratively merged and inverted (Wu, 1996). Zens and Ney (2003) exhaustively compare the IBM and ITG constraints, concluding that although the ITG constraints permit more flexible re-orderings, the IBM constraints result in higher BLEU scores.

Since our work falls under the IBM paradigm, we consider cost models for those constraints. We have said that linear distortion is the simplest cost model. The primary criticism of linear distortion is that it is unlexicalized, thus penalizing all re-orderings equally (Khalilov et al., 2009). When extended to phrases as in Equation (1), linear distortion is also agnostic to internal phrase alignments.

To remedy these deficiencies, Al-Onaizan and Papineni (2006) proposed a lexicalized, generative distortion model. Maximum likelihood estimates for inbound, outbound, and pairwise transitions are computed from automatic word alignments. But no estimate of future cost is included, and their model cannot easily accommodate features defined over the entire translation sequence. As for experimental results, they use a distortion limit that is half of what we report, and compare against a baseline that lacks a distortion model entirely. Neither their model nor ours requires generation of lattices prior to search (Zhang et al., 2007; Niehues and Kolss, 2009).

Lexicalized re-ordering models are the other significant approach to re-ordering. These models make local predictions about the next phrase to be translated during decoding, typically assigning costs to one of three categories: *monotone*, *swap*, or *discontinuous*. Both generative (Tillmann, 2004; Och and Ney, 2004; Koehn et al., 2007) and discriminative training (Tillmann and Zhang, 2005; Zens and Ney, 2006; Liang et al., 2006) algorithms have been proposed. Recently, Galley and Manning (2008) introduced a hierarchical model capable of analyzing alignments beyond adjacent phrases. Our discriminative distortion framework is not designed as a replacement for lexicalized re-ordering models, but as a substitute for linear distortion.

Finally, we comment on differences between our Arabic-English results and the well-known high distortion system of Zollmann et al. (2008), who find optimal baseline performance at a distortion limit of 9. First, they use approximately two orders of magnitude more training data, which allows them to extract much longer phrases (12 tokens v. our maximum of 7). In this setting, many Arabic-English reorderings can be captured in the phrase table. Second, their "Full" system uses three language models each trained with significantly more data than our single model. Finally, although they use a lexicalized re-ordering model, no details are given about the baseline distortion cost model.

7 Conclusion

We have presented a discriminative cost framework that both estimates future distortion cost and learns principled cost curves. The model delivers a statistically significant $+2.32$ BLEU improvement over Moses at a high distortion limit. Unlike previous discriminative local orientation models (Zens and Ney, 2006), our framework permits the definition of global features. The evaluation in this paper used context-free features to simplify the decoder integration, but we expect that context-dependent features could result in gains for other language pairs with more complex re-ordering phenomena.

Acknowledgements

We thank the three anonymous reviewers and Daniel Cer for constructive comments, and Claude Reichard for editorial assistance. The first author is supported by a National Defense Science and Engineering Graduate (NDSEG) fellowship. This paper is based on work supported in part 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

Y Al-Onaizan and K Papineni. 2006. Distortion models for statistical machine translation. In *ACL*.

G Andrew and J Gao. 2007. Scalable training of L1-regularized log-linear models. In *ICML*.

M Auli, A Lopez, H Hoang, and P Koehn. 2009. A systematic analysis of translation model search spaces. In *WMT*.

E Avramidis and P Koehn. 2008. Enriching morphologically poor languages for statistical machine translation. In *ACL*.

A Berger, P Brown, S Della Pietra, V Della Pietra, A Kehler, and R Mercer. 1996. Language translation apparatus and method using context-based translation models. *US Patent 5,510,981*.

D Cer, M Galley, D Jurafsky, and C D Manning. 2010. Phrasal: A statistical machine translation toolkit for exploring new model features. In *NAACL, Demonstration Session*.

S Chen and R Rosenfeld. 1999. A Gaussian prior for smoothing maximum entropy models. Technical Report CMU-CS-99-10S, Carnegie Mellon University.

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

M Collins, P Koehn, and I Kucerova. 2005. Clause restructuring for statistical machine translation. In *ACL*.

M Galley and C D Manning. 2008. A simple and effective hierarchical phrase reordering model. In *EMNLP*.

M Galley, S Green, D Cer, P-C Chang, and C D Manning. 2009. Stanford University's Arabic-to-English statistical machine translation system for the 2009 NIST evaluation. Technical report, Stanford University.

J Goodman. 2004. Exponential priors for maximum entropy models. In *NAACL*.

JH Greenberg, 1966. *Some universals of grammar with particular reference to the order of meaningful elements*, pages 73–113. London: MIT Press.

N Habash and F Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *NAACL*.

N Habash. 2007. Syntactic preprocessing for statistical machine translation. In *MT Summit XI*.

S Kanthak, D Vilar, E Matusov, R Zens, and H Ney. 2005. Novel reordering approaches in phrase-based statistical machine translation. In *ACL Workshop on Building and Using Parallel Texts*.

M Khalilov, J A R Fonollosa, and M Dras. 2009. Coupling hierarchical word reordering and decoding in phrase-based statistical machine translation. In *SSST*.

P Koehn, F J Och, and D Marcu. 2003. Statistical phrase-based translation. In *NAACL*.

P Koehn, H Hoang, A Birch, C Callison-Burch, M Federico, N Bertoldi, B Cowan, W Shen, C Moran, R Zens, C Dyer, O Bojar, A Constantin, and E Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL, Demonstration Session*.

P Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*.

P Liang, B Taskar, and D Klein. 2006. Alignment by agreement. In *NAACL*.

Y Marton and P Resnik. 2008. Soft syntactic constraints for hierarchical phrased-based translation. In *ACL*.

R C Moore and C Quirk. 2007. Faster beam-search decoding for phrasal statistical machine translation. In *MT Summit XI*.

J Niehues and M Kolss. 2009. A POS-based model for long-range reorderings in SMT. In *WMT*.

F J Och and H Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449.

F J Och. 2003. Minimum error rate training for statistical machine translation. In *ACL*.

K Papineni, S Roukos, T Ward, and W-J Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *ACL*.

S Riezler and J T Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing in MT. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization (MTSE'05)*.

C Tillmann and T Zhang. 2005. A localized prediction model for statistical machine translation. In *ACL*.

C Tillmann. 2004. A unigram orientation model for statistical machine translation. In *NAACL*.

K Toutanova, D Klein, C D Manning, and Y Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.

D Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *ACL*.

R Zens and H Ney. 2003. A comparative study on reordering constraints in statistical machine translation. In *ACL*.

R Zens and H Ney. 2006. Discriminative reordering models for statistical machine translation. In *WMT*.

R Zens. 2008. *Phrase-based Statistical Machine Translation: Models, Search, Training*. Ph.D. thesis, RWTH Aachen University.

Y Zhang, R Zens, and H Ney. 2007. Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *SSST*.

A Zollmann, A Venugopal, F J Och, and J Ponte. 2008. A systematic comparison of phrase-based, hierarchical and syntax-augmented statistical MT. In *COLING*.