Distortion Models For Statistical Machine Translation

Yaser Al-Onaizan and Kishore Papineni
IBM T.J. Watson Research Center
1101 Kitchawan Road
Yorktown Heights, NY 10598, USA
{onaizan, papineni}@us.ibm.com

Abstract

In this paper, we argue that n-gram language models are not sufficient to address word reordering required for Machine Translation. We propose a new distortion model that can be used with existing phrase-based SMT decoders to address those n-gram language model limitations. We present empirical results in Arabic to English Machine Translation that show statistically significant improvements when our proposed model is used. We also propose a novel metric to measure word order similarity (or difference) between any pair of languages based on word alignments.

1 Introduction

A language model is a statistical model that gives a probability distribution over possible sequences of words. It computes the probability of producing a given word w_1 given all the words that precede it in the sentence. An n-gram language model is an n-th order Markov model where the probability of generating a given word depends only on the last $n-1$ words immediately preceding it and is given by the following equation:

$$P(w_1^k) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1^{n-1}) \quad (1)$$

where $k >= n$.

N-gram language models have been successfully used in Automatic Speech Recognition (ASR) as was first proposed by (Bahl et al., 1983). They play an important role in selecting among several candidate word realization of a given acoustic signal. N-gram language models have also been used in Statistical Machine Translation (SMT) as proposed by (Brown et al., 1990; Brown et al., 1993). The run-time search procedure used to find the most likely translation (or transcription in the case of Speech Recognition) is typically referred to as *decoding*.

There is a fundamental difference between decoding for machine translation and decoding for speech recognition. When decoding a speech signal, words are generated in the same order in which their corresponding acoustic signal is consumed. However, that is not necessarily the case in MT due to the fact that different languages have different word order requirements. For example, in Spanish and Arabic adjectives are mainly noun post-modifiers, whereas in English adjectives are noun pre-modifiers. Therefore, when translating between Spanish and English, words must usually be reordered.

Existing statistical machine translation decoders have mostly relied on language models to select the proper word order among many possible choices when translating between two languages. In this paper, we argue that a language model is not sufficient to adequately address this issue, especially when translating between languages that have very different word orders as suggested by our experimental results in Section 5. We propose a new distortion model that can be used as an additional component in SMT decoders. This new model leads to significant improvements in MT quality as measured by BLEU (Papineni et al., 2002). The experimental results we report in this paper are for Arabic-English machine translation of news stories.

We also present a novel method for measuring word order similarity (or differences) between any given pair of languages based on word alignments as described in Section 3.

The rest of this paper is organized as follows. Section 2 presents a review of related work. In Section 3 we propose a method for measuring the distortion between any given pair of languages. In Section 4, we present our proposed distortion model. In Section 5, we present some empirical results that show the utility of our distortion model for statistical machine translation systems. Then, we conclude this paper with a discussion in Section 6.

2 Related Work

Different languages have different word order requirements. SMT decoders attempt to generate translations in the proper word order by attempting many possible

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 529–536,
Sydney, July 2006.

word reorderings during the translation process. Trying all possible word reordering is an NP-Complete problem as shown in (Knight, 1999), which makes searching for the optimal solution among all possible permutations computationally intractable. Therefore, SMT decoders typically limit the number of permutations considered for efficiency reasons by placing reordering restrictions. Reordering restrictions for word-based SMT decoders were introduced by (Berger et al., 1996) and (Wu, 1996). (Berger et al., 1996) allow only reordering of at most n words at any given time. (Wu, 1996) propose using contiguity restrictions on the reordering. For a comparison and a more detailed discussion of the two approaches see (Zens and Ney, 2003).

A different approach to allow for a limited reordering is to reorder the input sentence such that the source and the target sentences have similar word order and then proceed to monotonically decode the reordered source sentence.

Monotone decoding translates words in the same order they appear in the source language. Hence, the input and output sentences have the same word order. Monotone decoding is very efficient since the optimal decoding can be found in polynomial time. (Tillmann et al., 1997) proposed a DP-based monotone search algorithm for SMT. Their proposed solution to address the necessary word reordering is to rewrite the input sentence such that it has a similar word order to the desired target sentence. The paper suggests that reordering the input reduces the translation error rate. However, it does not provide a methodology on how to perform this reordering.

(Xia and McCord, 2004) propose a method to automatically acquire rewrite patterns that can be applied to any given input sentence so that the rewritten source and target sentences have similar word order. These rewrite patterns are automatically extracted by parsing the source and target sides of the training parallel corpus. Their approach show a statistically-significant improvement over a phrase-based monotone decoder. Their experiments also suggest that allowing the decoder to consider some word order permutations *in addition to* the rewrite patterns already applied to the source sentence actually decreases the BLEU score.

Rewriting the input sentence whether using syntactic rules or heuristics makes hard decisions that can not be undone by the decoder. Hence, reordering is better handled during the search algorithm and as part of the optimization function.

Phrase-based monotone decoding does not directly address word order issues. Indirectly, however, the phrase dictionary[1] in phrase-based decoders typically captures local reorderings that were seen in the training data. However, it fails to generalize to word reorderings that were never seen in the training data. For example, a phrase-based decoder might translate the Arabic phrase *AlwlAyAt AlmtHdp*[2] correctly into English as *the United States* if it was seen in its training data, was aligned correctly, and was added to the phrase dictionary. However, if the phrase *Almmlkp AlmtHdp* is not in the phrase dictionary, it will not be translated correctly by a monotone phrase decoder even if the individual units of the phrase *Almmlkp* and *AlmtHdp*, and their translations (*Kingdom* and *United*, respectively) are in the phrase dictionary since that would require swapping the order of the two words.

(Och et al., 1999; Tillmann and Ney, 2003) relax the monotonicity restriction in their phrase-based decoder by allowing a restricted set of word reorderings. For their translation task, word reordering is done only for words belonging to the verb group. The context in which they report their results is a Speech-to-Speech translation from German to English.

(Yamada and Knight, 2002) propose a syntax-based decoder that restrict word reordering based on reordering operations on syntactic parse-trees of the input sentence. They reported results that are better than word-based IBM4-like decoder. However, their decoder is outperformed by phrase-based decoders such as (Koehn, 2004), (Och et al., 1999), and (Tillmann and Ney, 2003) . Phrase-based SMT decoders mostly rely on the language model to select among possible word order choices. However, in our experiments we show that the language model is not reliable enough to make the choices that lead to a better MT quality. This observation is also reported by (Xia and McCord, 2004).We argue that the distortion model we propose leads to a better translation as measured by BLEU.

Distortion models were first proposed by (Brown et al., 1993) in the so-called IBM Models. IBM Models 2 and 3 define the distortion parameters in terms of the word positions in the sentence pair, not the actual words at those positions. Distortion probability is also conditioned on the source and target sentence lengths. These models do not generalize well since their parameters are tied to absolute word position within sentences which tend to be different for the same words across sentences. IBM Models 4 and 5 alleviate this limitation by replacing absolute word positions with relative positions. The latter models define the distortion parameters for a cept (one or more words). This models phrasal movement better since words tend to move in blocks and not independently. The distortion is conditioned on classes of the aligned source and target words. The entire source and target vocabularies are reduced to a small number of classes (e.g., 50) for the purpose of estimating those parameters.

Similarly, (Koehn et al., 2003) propose a relative distortion model to be used with a phrase decoder. The model is defined in terms of the difference between the position of the current phrase and the position of the previous phrase in the source sentence. It does not con-

[1] Also referred to in the literature as the set of blocks or clumps.

[2] Arabic text appears throughout this paper in Tim Buckwalter's Romanization.

Arabic	Ezp_1 AbrAhym_2 ystqbl_3 $\text{ms\&wlA}_4$ AqtSAdyA_5 sEwdyA_6 fy_7 bgdAd_8
English	Izzet_1 Ibrahim_2 Meets_3 Saudi_4 Trade_5 official_6 in_7 Baghdad_8
Word Alignment	(Ezp_1,Izzet_1) (AbrAhym_2,Ibrahim_2) (ystqbl_3,Meets_3) ($\text{ms\&wlA}_4$,official_6) (AqtSAdyA_5,Trade_5) (sEwdyA_6,Saudi_4) (fy_7,in_7) (bgdAd_8,Baghdad_8)
Reordered English	Izzet_1 Ibrahim_2 Meets_3 official_6 Trade_5 Saudi_4 in_7 Baghdad_8

Table 1: Alignment-based word reordering. The indices are not part of the sentence pair, they are only used to illustrate word positions in the sentence. The indices in the reordered English denote word position in the original English order.

sider the words in those positions.

The distortion model we propose assigns a probability distribution over possible relative jumps conditioned on source words. Conditioning on the source words allows for a much more fine-grained model. For instance, words that tend to act as modifers (e.g., adjectives) would have a different distribution than verbs or nouns. Our model's parameters are directly estimated from word alignments as we will further explain in Section 4. We will also show how to generalize this word distortion model to a phrase-based model.

(Och et al., 2004; Tillman, 2004) propose orientation-based distortion models lexicalized on the phrase level. There are two important distinctions between their models and ours. First, they lexicalize their model on the phrases, which have many more parameters and hence would require much more data to estimate reliably. Second, their models consider only the direction (i.e., orientation) and not the relative jump.

We are not aware of any work on measuring word order differences between a given language pair in the context of statistical machine translation.

3 Measuring Word Order Similarity Between Two Language

In this section, we propose a simple, novel method for measuring word order similarity (or differences) between any given language pair. This method is based on word-alignments and the BLEU metric.

We assume that we have word-alignments for a set of sentence pairs. We first reorder words in the target sentence (e.g., English when translating from Arabic to English) according to the order in which they are aligned to the source words as shown in Table 1. If a target word is not aligned, then, we assume that it is aligned to the same source word that the preceding aligned target word is aligned to.

Once the reordered target (here English) sentences are generated, we measure the distortion between the language pair by computing the BLEU[3] score between the original target and reordered target, treating the original target as the reference.

Table 2 shows these scores for Arabic-English and Chinese-English. The word alignments we use are both annotated manually by human annotators. The Arabic-English test set is the NIST MT Evaluation 2003 test set. It contains 663 segments (i.e., sentences). The Arabic side consists of 16,652 tokens and the English consists of 19,908 tokens. The Chinese-English test set contains 260 segments. The Chinese side is word segmented and consists of 4,319 tokens and the English consists of 5,525 tokens.

As suggested by the BLEU scores reported in Table 2, Arabic-English has more word order differences than Chinese-English. The difference in n-gPrec is bigger for smaller values of n, which suggests that Arabic-English has more local word order differences than in Chinese-English.

4 Proposed Distortion Model

The distortion model we are proposing consists of three components: **outbound**, **inbound**, and **pair** distortion. Intuitively our distortion models attempt to capture the order in which source words need to be translated. For instance, the outbound distortion component attempts to capture what is typically translated immediately after the word that has just been translated. Do we tend to translate words that precede it or succeed it? Which word position to translate next?

Our distortion parameters are directly estimated from word alignments by simple counting over alignment links in the training data. Any aligner such as (Al-Onaizan et al., 1999) or (Vogel et al., 1996) can be used to obtain word alignments. For the results reported in this paper word alignments were obtained using a maximum-posterior word aligner[4] described in (Ge, 2004).

We will illustrate the components of our model with a partial word alignment. Let us assume that our source sentence[5] is $(f_{10}, f_{250}, f_{300})$[6], and our target sentence is (e_{410}, e_{20}), and their word alignment is $a = ((f_{10}, e_{410}), (f_{300}, e_{20}))$. Word Alignment a can

[3]the BLEU scores reported throughout this paper are for case-sensitive BLEU. The number of references used is also reported (e.g., BLEUr1n4c: $r1$ means 1 reference, $n4$ means upto 4-gram are considred, c means case sensitive).

[4]We also estimated distortion parameters using a Maximum Entropy aligner and the differences were negligible.

[5]In practice, we add special symbols at the start and end of the source and target sentences, we also assume that the start symbols in the source and target are aligned, and similarly for the end symbols. Those special symbols are omitted in our example for ease of presentation.

[6]The indices here represent source and target vocabulary ids.

N-gram Precision	Arabic-English	Chinese-English
1-gPrec	1	1
2-gPrec	0.6192	0.7378
3-gPrec	0.4547	0.5382
4-gPrec	0.3535	0.3990
5-gPrec	0.2878	0.3075
6-gPrec	0.2378	0.2406
7-gPrec	0.1977	0.1930
8-gPrec	0.1653	0.1614
9-gPrec	0.1380	0.1416
BLEUr1n4c	0.3152	0.3340
95% Confidence σ	0.0180	0.0370

Table 2: Word order similarity for two language pairs: Arabic-English and Chinese-English. n-gPrec is the n-gram precision as defined in BLEU.

be rewritten as $a_1 = 1$ and $a_2 = 3$ (i.e., the second target word is aligned to the third source word). From this partial alignment we increase the counts for the following outbound, inbound, and pair distortions: $P_o(\delta = +2|f_{10})$, $P_i(\delta = +2|f_{300})$. and $P_p(\delta = +2|f_{10}, f_{300})$.

Formally, our distortion model components are defined as follows:

Outbound Distortion:

$$P_o(\delta|f_i) = \frac{C(\delta|f_i)}{\sum_k C(\delta_k\ |f_i)} \quad (2)$$

where f_i is a foreign word (i.e., Arabic in our case), δ is the step size, and $C(\delta|f_i)$ is the observed count of this parameter over all word alignments in the training data. The value for δ, in theory, ranges from $-max$ to $+max$ (where max is the maximum source sentence length observed), but in practice only a small number of those step sizes are observed in the training data, and hence, have non-zero value).

Inbound Distortion:

$$P_i(\delta|f_j) = \frac{C(\delta|f_j)}{\sum_k C(\delta_k|f_j)} \quad (3)$$

Pairwise Distortion:

$$P_p(\delta|f_i, f_j) = \frac{C(\delta|f_i, f_j)}{\sum_k C(\delta_k|f_i, f_j)} \quad (4)$$

In order to use these probability distributions in our decoder, they are then turned into costs. The outbound distortion cost is defined as:

$$C_o(\delta|f_i) = \log\{\alpha P_o(\delta|f_i) + (1-\alpha)P_s(\delta)\} \quad (5)$$

where $P_s(\delta)$ is a smoothing distribution [7] and α is a linear-mixture parameter [8].

[7] The smoothing we use is a geometrically decreasing distribution as the step size increases.

[8] For the experiments reported here we use $\alpha = 0.1$, which is set empirically.

The inbound and pair costs ($C_i(\delta|f_i)$ and $C_p(\delta|f_i, f_j)$) can be defined in a similar fashion.

So far, our distortion cost is defined in terms of words, not phrases. Therefore, we need to generalize the distortion cost in order to use it in a phrase-based decoder. This generalization is defined in terms of the internal word alignment within phrases (we used the Viterbi word alignment). We illustrate this with an example: Suppose the last position translated in the source sentence so far is n and we are to cover a source phrase *p=wlAyp wA$nTn* that begins at position m in the source sentence. Also, suppose that our phrase dictionary provided the translation *Washington State*, with internal word alignment $a = (a_1 = 2, a_2 = 1)$ (i.e., *a=(<Washington,wA$nTn>, <State,wlAyp>)*, then the outbound phrase cost is defined as:

$$C_o(p, n, m, a) = C_o(\delta = (m-n)|f_n) + \sum_{i=1}^{l-1} C_o(\delta = (a_{i+1} - a_i)\ |f_{a_i}) \quad (6)$$

where l is the length of the target phrase, a is the internal word alignment, f_n is source word at position n (in the sentence), and f_{a_i} is the source word that is aligned to the i-th word in the target side of the phrase (not the sentence).

The inbound and pair distortion costs (i..e, $C_i(p, n, m, a)$ and $C_p(p, n, m, a)$) can be defined in a similar fashion.

The above distortion costs are used in conjunction with other cost components used in our decoder. The ultimate word order choice made is influenced by both the language model cost as well as the distortion cost.

5 Experimental Results

The phrase-based decoder we use is inspired by the decoder described in (Tillmann and Ney, 2003) and similar to that described in (Koehn, 2004). It is a multi-stack, multi-beam search decoder with n stacks (where n is the length of the source sentence being decoded)

s	0	1	1	1	1	1	2	2	2	2
w	0	4	6	8	10	12	4	6	8	10
BLEUr1n4c	0.5617	0.6507	0.6443	0.6430	0.6461	0.6456	0.6831	0.6706	0.6609	0.6596

2	3	3	3	3	3	4	4	4	4	4
12	4	6	8	10	12	4	6	8	10	12
0.6626	**0.6919**	0.6751	0.6580	0.6505	0.6490	0.6851	0.6592	0.6317	0.6237	**0.6081**

Table 3: BLEU scores for the word order restoration task. The BLEU scores reported here are with 1 reference. The input is the reordered English in the reference. The 95% Confidence σ ranges from 0.011 to 0.016

and a beam associated with each stack as described in (Al-Onaizan, 2005). The search is done in n time steps. In time step i, only hypotheses that cover exactly i source words are extended. The beam search algorithm attempts to find the translation (i.e., hypothesis that covers all source words) with the minimum cost as in (Tillmann and Ney, 2003) and (Koehn, 2004) . The distortion cost is added to the log-linear mixture of the hypothesis extension in a fashion similar to the language model cost.

A hypothesis covers a subset of the source words. The final translation is a hypothesis that covers all source words and has the minimum cost among all possible [9] hypotheses that cover all source words. A hypothesis h is extended by matching the phrase dictionary against source word sequences in the input sentence that are not covered in h. The cost of the new hypothesis $C(h_{new}) = C(h) + C(e)$, where $C(e)$ is the cost of this extension. The main components of the cost of extension e can be defined by the following equation:

$$C(e) = \lambda_1 C_{LM}(e) + \lambda_2 C_{TM}(e) + \lambda_3 C_D(e)$$

where $C_{LM}(e)$ is the language model cost, $C_{TM}(e)$ is the translation model cost, and $C_D(e)$ is the distortion cost. The extension cost depends on the hypothesis being extended, the phrase being used in the extension, and the source word positions being covered.

The word reorderings that are explored by the search algorithm are controlled by two parameters s and w as described in (Tillmann and Ney, 2003). The first parameter s denotes the number of source words that are temporarily skipped (i.e., temporarily left uncovered) during the search to cover a source word to the right of the skipped words. The second parameter is the window width w, which is defined as the distance (in number of source words) between the left-most uncovered source word and the right-most covered source word.

To illustrate these restrictions, let us assume the input sentence consists of the following sequence (f_1, f_2, f_3, f_4). For s=1 and w=2, the permissible permutations are (f_1, f_2, f_3, f_4), (f_2, f_1, f_3, f_4), (f_2, f_3, f_1, f_4), (f_1, f_3, f_2, f_4),(f_1, f_3, f_4, f_2), and (f_1, f_2, f_4, f_3).

[9] Exploring all possible hypothesis with all possible word permutations is computationally intractable. Therefore, the search algorithm gives an approximation to the optimal solution. *All possible hypotheses* refers to all hypotheses that were explored by the decoder.

5.1 Experimental Setup

The experiments reported in this section are in the context of SMT from Arabic into English. The training data is a 500K sentence-pairs subsample of the 2005 Large Track Arabic-English Data for NIST MT Evaluation.

The language model used is an interpolated trigram model described in (Bahl et al., 1983). The language model is trained on the LDC English GigaWord Corpus.

The test set used in the experiments in this section is the 2003 NIST MT Evaluation test set (which is not part of the training data).

5.2 Reordering with Perfect Translations

In the experiments in this section, we show the utility of a trigram language model in restoring the correct word order for English. The task is a simplified translation task, where the input is reordered English (English written in Arabic word order) and the output is English in the correct order. The source sentence is a reordered English sentence in the same manner we described in Section 3. The objective of the decoder is to recover the correct English order.

We use the same phrase-based decoder we use for our SMT experiments, except that only the language model cost is used here. Also, the phrase dictionary used is a one-to-one function that maps every English word in our vocabulary to itself. The language model we use for the experiments reported here is the same as the one used for other experiments reported in this paper.

The results in Table 3 illustrate how the language model performs reasonably well for local reorderings (e.g., for $s = 3$ and $w = 4$), but its perfromance deteriorates as we relax the reordering restrictions by increasing the reordering window size (w).

Table 4 shows some examples of original English, English in Arabic order, and the decoder output for two different sets of reordering parameters.

5.3 SMT Experiments

The phrases in the phrase dictionary we use in the experiments reported here are a combination

Eng_Ar	Opposition Iraqi Prepares for Meeting mid - January in Kurdistan
Orig. Eng.	Iraqi Opposition Prepares for mid - January Meeting in Kurdistan
Output1	Iraqi Opposition Meeting Prepares for mid - January in Kurdistan
Output2	Opposition Meeting Prepares for Iraqi Kurdistan in mid - January
Eng_Ar	Head of Congress National Iraqi Visits Kurdistan Iraqi
Orig. Eng.	Head of Iraqi National Congress Visits Iraqi Kurdistan
Output1	Head of Iraqi National Congress Visits Iraqi Kurdistan
Output2	Head Visits Iraqi National Congress of Iraqi Kurdistan
Eng_Ar	House White Confirms Presence of Tape New Bin Laden
Orig. Eng.	White House Confirms Presence of New Bin Laden Tape
Output1	White House Confirms Presence of Bin Laden Tape New
Output2	White House of Bin Laden Tape Confirms Presence New

Table 4: Examples of reordering with perfect translations. The examples show English in Arabic order (Eng_Ar.), English in its original order (Orig. Eng.) and decoding with two different parameter settings. Output1 is decoding with (s=3,w=4). Output2 is decoding with (s=4,w=12). The sentence lengths of the examples presented here are much shorter than the average in our test set (~ 28.5).

s	w	Distortion Used?	BLEUr4n4c
0	0	NO	**0.4468**
1	8	NO	0.4346
1	8	YES	0.4715
2	8	NO	0.4309
2	8	YES	0.4775
3	8	NO	0.4283
3	8	YES	**0.4792**
4	8	NO	0.4104
4	8	YES	0.4782

Table 5: BLEU scores for the Arabic-English machine translation task. The 95% Confidence σ ranges from 0.0158 to 0.0176. s is the number of words temporarily skipped, and w is the word permutation window size.

of phrases automatically extracted from maximum-posterior alignments and maximum entropy alignments. Only phrases that conform to the so-called consistent alignment restrictions (Och et al., 1999) are extracted.

Table 5 shows BLEU scores for our SMT decoder with different parameter settings for skip s, window width w, with and without our distortion model. The BLEU scores reported in this table are based on 4 reference translations. The language model, phrase dictionary, and other decoder tuning parameters remain the same in all experiments reported in this table.

Table 5 clearly shows that as we open the search and consider wider range of word reorderings, the BLEU score decreases in the absence of our distortion model when we rely solely on the language model. Wrong reorderings look attractive to the decoder via the language model which suggests that we need a richer model with more parameter. In the absence of richer models such as the proposed distortion model, our results suggest that it is best to decode monotonically and only allow local reorderings that are captured in our phrase dictionary.

However, when the distortion model is used, we see statistically significant increases in the BLEU score as we consider more word reorderings. The best BLEU score achieved when using the distortion model is 0.4792 , compared to a best BLEU score of 0.4468 when the distortion model is not used.

Our results on the 2004 and 2005 NIST MT Evaluation test sets using the distortion model are 0.4497 and 0.4646[10], respectively.

Table 6 shows some Arabic-English translation examples using our decoder with and without the distortion model.

6 Conclusion and Future Work

We presented a new distortion model that can be integrated with existing phrase-based SMT decoders. The proposed model shows statistically significant improvement over a state-of-the-art phrase-based SMT decoder. We also showed that n-gram language mod-

[10]The MT05 BLEU score is the from the official NIST evaluation. The MT04 BLEU score is only our second run on MT04.

Input (Ar)	kwryA Al$mAlyp mstEdp llsmAH lwA$nTn bAltHqq mn AnhA lA tSnE AslHp nwwyp
Ref. (En)	North Korea Prepared to allow Washington to check it is not Manufacturing Nuclear Weapons
Out1	North Korea to Verify Washington That It Was Not Prepared to Make Nuclear Weapons
Out2	North Korea Is Willing to Allow Washington to Verify It Does Not Make Nuclear Weapons
Input (Ar)	wAkd Aldblwmasy An ”AnsHAb (kwryA Al$mAlyp mn AlmEAhdp) ybd> AEtbArA mn Alywm”.
Ref. (En)	The diplomat confirmed that ”North Korea’s withdrawal from the treaty starts as of today.”
Out1	The diplomat said that ” the withdrawal of the Treaty (start) North Korea as of today. ”
Out2	The diplomat said that the ” withdrawal of (North Korea of the treaty) will start as of today ”.
Input (Ar)	snrfE *lk AmAm Almjls Aldstwry”.
Ref. (En)	We will bring this before the Constitutional Assembly.”
Out1	The Constitutional Council to lift it. ”
Out2	This lift before the Constitutional Council ”.
Input (Ar)	wAkd AlbrAdEy An mjls AlAmn ”ytfhm” An 27 kAnwn AlvAny/ynAyr lys mhlp nhA}yp.
Ref. (En)	Baradei stressed that the Security Council ”appreciates” that January 27 is not a final ultimatum.
Out1	Elbaradei said that the Security Council ” understand ” that is not a final period January 27.
Out2	Elbaradei said that the Security Council ” understand ” that 27 January is not a final period.

Table 6: Selected examples of our Arabic-English SMT output. The English is one of the human reference translations. Output 1 is decoding without the distortion model and (s=4, w=8), which corresponds to 0.4104 BLEU score. Output 2 is decoding with the distortion model and (s=3, w=8), which corresponds to 0.4792 BLEU score. The sentences presented here are much shorter than the average in our test set. The average length of the arabic sentence in the MT03 test set is ~ 24.7.

els are not sufficient to model word movement in translation. Our proposed distortion model addresses this weakness of the n-gram language model.

We also propose a novel metric to measure word order similarity (or differences) between any pair of languages based on word alignments. Our metric shows that Chinese-English have a closer word order than Arabic-English.

Our proposed distortion model relies solely on word alignments and is conditioned on the source words. The majority of word movement in translation is mainly due to syntactic differences between the source and target language. For example, Arabic is verb-initial for the most part. So, when translating into English, one needs to move the verb after the subject, which is often a long compounded phrase. Therefore, we would like to incorporate syntactic or part-of-speech information in our distortion model.

Acknowledgment

This work was partially supported by DARPA GALE program under contract number HR0011-06-2-0001. It was also partially supported by DARPA TIDES program monitored by SPAWAR under contract number N66001-99-2-8916.

References

Yaser Al-Onaizan, Jan Curin, Michael Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz-Josef Och, David Purdy, Noah Smith, and David Yarowsky. 1999. Statistical Machine Translation: Final Report, Johns Hopkins University Summer Workshop (WS 99) on Language Engineering, Center for Language and Speech Processing, Baltimore, MD.

Yaser Al-Onaizan. 2005. IBM Arabic-to-English MT Submission. *Presentation given at DARPA/TIDES NIST MT Evaluation workshop*.

Lalit R. Bahl, Frederick Jelinek, and Robert L. Mercer. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2):179–190.

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Andrew S. Kehler, and Robert L. Mercer. 1996. Language Translation Apparatus and Method of Using Context-Based Translation Models. *United States Patent, Patent Number 5510981*, April.

Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Niyu Ge. 2004. Improvements in Word Alignments. *Presentation given at DARPA/TIDES NIST MT Evaluation workshop*.

Kevin Knight. 1999. Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, 25(4):607–615.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Main Proceedings*, pages 127–133, Edmonton, Alberta, Canada, May 27 – June 1. Association for Computational Linguistics.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, pages 115–124, Washington DC, September-October. The Association for Machine Translation in the Americas (AMTA).

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Joint Conf. of Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, College Park, Maryland.

Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A Smorgasbord of Features for Statistical Machine Translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 161–168, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of machine translation. In *40th Annual Meeting of the Association for Computational Linguistics (ACL 02)*, pages 311–318, Philadelphia, PA, July.

Christoph Tillman. 2004. A unigram orientation model for statistical machine translation. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *HLT-NAACL 2004: Short Papers*, pages 101–104, Boston, Massachusetts, USA, May 2 - May 7. Association for Computational Linguistics.

Christoph Tillmann and Hermann Ney. 2003. Word Re-ordering and a DP Beam Search Algorithm for Statistical Machine Translation. *Computational Linguistics*, 29(1):97–133.

Christoph Tillmann, Stephan Vogel, Hermann Ney, and Alex Zubiaga. 1997. A DP-Based Search Using Monotone Alignments in Statistical Translation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 289–296, Madrid. Association for Computational Linguistics.

Stefan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-BasedWord Alignment in Statistical Machine Translation. In *Proc. of the 16th Int. Conf. on Computational Linguistics (COLING 1996)*, pages 836–841, Copenhagen, Denmark, August.

Dekai Wu. 1996. A Polynomial-Time Algorithm for Statistical Machine Translation. In *Proc. of the 34th Annual Conf. of the Association for Computational Linguistics (ACL 96)*, pages 152–158, Santa Cruz, CA, June.

Fei Xia and Michael McCord. 2004. Improving a Statistical MT System with Automatically Learned Rewrite Patterns. In *Proc. of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Kenji Yamada and Kevin Knight. 2002. A Decoder for Syntax-based Statistical MT. In *Proc. of the 40th Annual Conf. of the Association for Computational Linguistics (ACL 02)*, pages 303–310, Philadelphia, PA, July.

Richard Zens and Hermann Ney. 2003. A Comparative Study on Reordering Constraints in Statistical Machine Translation. In Erhard Hinrichs and Dan Roth, editors, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 144–151, Sapporo, Japan.