Improving Reordering in Statistical Machine Translation from Farsi

Evgeny Matusov
AppTek Inc.
Aachen, Germany
ematusov@apptek.com

Selçuk Köprü
AppTek Inc.
METU Technopolis
Ankara, Turkey
skopru@apptek.com

Abstract

In this paper, we propose a novel model for scoring reordering in phrase-based statistical machine translation (SMT) and successfully use it for translation from Farsi into English and Arabic. The model replaces the distance-based distortion model that is widely used in most SMT systems. The main idea of the model is to penalize each new deviation from the monotonic translation path. We also propose a way for combining this model with manually created reordering rules for Farsi which try to alleviate the difference in sentence structure between Farsi and English/Arabic by changing the position of the verb. The rules are used in the SMT search as soft constraints.

In the experiments on two general-domain translation tasks, the proposed penalty-based model improves the BLEU score by up to 1.5% absolute as compared to the baseline of monotonic translation, and up to 1.2% as compared to using the distance-based distortion model.

1 Introduction

In recent years, phrase-based SMT systems have achieved good translation quality. Yet one of the problems which these systems are often not able to solve is how to make the translation fluent, i.e. translate phrases in the right order. Usually, the basic costs for reordering a phrase in the SMT search are linear in the distance (Koehn, 2004). More complex models have been introduced, but in most cases they extend the simple distance-based distortion model.

In this work, we argue that the distance-based model has a number of disadvantages for SMT. We replace it with a novel reordering model. This model assigns a penalty for each new *run*, which is a new deviation from the monotonic translation path. We show experimentally that the proposed model outperforms the distance-based model.

Another focus of this paper is the combination of the run-based penalty model with hand-crafted reordering rules for Farsi. Applying manual reordering rules which rely on part-of-speech (POS) tags or syntactic parses in a preprocessing step often does not lead to improved translation quality, since the rules do not handle all exceptions, or fail because of erroneous parses. Therefore, we introduce such rules into the SMT search as soft constraints only, so that SMT hypotheses which follow the permutation of the source sentence defined by the rules are given a bonus.

This work is structured as follows. In Section 2, we review the related research on reordering in statistical MT, as well as on translation from Farsi in general. In Section 3, we first describe our baseline SMT system and then focus on the novel reordering models we propose. Section 4 describes how the hand-crafted reordering rules, which are used as soft constraints in the SMT, have been designed. Information on the Farsi part-of-speech tagger and the parser is also included. Section 5 presents the experimental results. It is followed by our conclusions.

2 Related work

The distance-based penalty model is used in many statistical phrase-based decoders, including the open-source decoder MOSES (Koehn et al., 2007). In (Zens and Ney, 2006), additionally a maximum-entropy reordering model is used to predict the ori-

entation of a phrase (left or right of the previously translated phrase). The words or word classes in the concerned phrase pairs are utilized as model features. Similar orientation or lexicalized reordering models have been proposed also in (Nagata et al., 2006; Koehn et al., 2005; Al-Onaizan and Papineni, 2006). However, they were almost always used in combination with the distance-based model. Other researchers tried to include reordering rules which either have been defined manually (Wang et al., 2007), or have been learned statistically from the reordering patterns in the parallel training data (Chen et al., 2006). In many cases the rules utilize POS tags (Rottmann and Vogel, 2007) or parses (Collins et al., 2005). The rules were either applied before SMT, or defined a reordering search space for SMT by representing reordering alternatives in a word graph (Zhang et al., 2007; Li et al., 2007).

Previous work on Farsi MT is limited. The rule-based MT system described in (Amtrup et al., 2000) is one of the first systems that translates from Farsi to English. The authors introduce a toolkit and its application to Farsi. Saedi et al. (2009) gives an overview of a bidirectional English - Farsi MT system containing rule-based, knowledge-based and corpus-based components. In (Deng and Zhou, 2009), the authors present a method for word alignment symmetrization and combination and test on a Farsi-to-English task. Several small vocabulary Farsi-to-English systems have been developed within the TransTac project for real-time dialogue applications (Kathol and Zheng, 2008; Kao et al., 2008). We are not aware of any work on Farsi-to-Arabic MT.

3 Reordering in phrase-based statistical MT

3.1 Baseline MT system

The baseline MT system is a state-of-the-art phrase-based translation system similar to (Koehn et al., 2007) and (Zens, 2008). In this system, a target language translation $e_1^I = e_1 \dots e_i \dots e_I$ for the source language sentence $f_1^J = f_1 \dots f_j \dots f_J$ is found by maximizing the posterior probability $Pr(e_1^I|f_1^J)$. This probability is modeled directly using a log-linear combination of several models. The best translation is found with the following decision rule:

$$\hat{e}_1^{\hat{I}} = \underset{I,e_1^I}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(e_1^I, f_1^J) \right\} \quad (1)$$

The model scaling factors λ_m for the features h_m are trained with respect to the final translation quality measured by an error criterion (Och, 2003). The baseline system includes an n-gram language model, a phrase translation model, and a word-based lexicon model as the main features. The latter two models are used in both directions: $p(f|e)$ and $p(e|f)$. Further log-linear model features include a word penalty and a phrase penalty. Finally, reordering models are also used as features.

The phrase-based search consists of two parts. First, those contiguous phrases in the source sentence are identified which have translation candidates in the phrase table. This phrase matching is done efficiently using an algorithm based on the work of (Zens, 2008). The second phase is the source cardinality-synchronous search (SCSS) implemented with dynamic programming. The goal of the search is to find the most probable segmentation of the source sentence into K non-empty non-overlapping contiguous blocks with boundaries (b_k, j_k), $1 \leq b_k \leq j_k \leq J$; $1 \leq k \leq K$, select the most probable permutation of those blocks, and choose the best phrasal translations for each of the blocks at the same time. The concatenation of the translations of the permuted blocks yields a translation of the whole sentence.

The search algorithm proceeds synchronously with the cardinality k of the already translated source positions. With each partial hypothesis a coverage set $C \subseteq \{1, \dots, J\}$ is associated, it holds $k = |C|$. Given a hypothesis with cardinality k, the decoder selects a range of source positions $b_k, \dots, j_k$ for which there is no overlap with the already translated positions, i. e. $C \cap \{b_k, \dots, j_k\} = \emptyset$. Then, with each target phrase translation of this source range, the current hypothesis is extended, and a new hypothesis with cardinality $k' = |C \cup \{b_k, \dots, j_k\}|$ is formed. This hypothesis will be processed in step k' of the algorithm.

The reordering in the SCSS is performed through selection of, in principle, arbitrary source position ranges in each step of the algorithm. However, this selection is usually limited by some reordering constraints in order to reduce computational complexity and/or model the reordering with automatically

learned or linguistically motivated rules. The possible choice of constraints is described in the next section.

3.2 Reordering constraints

In practice, coverage sets are implemented using bitvectors $\bar{c}$ of dimension J. For any source position j, it holds $\bar{c}_j = 1$ if and only if $j \in C$. Representation with bitvectors makes the definition of reordering constraints more convenient.

A trivial constraint that makes the search process monotonic is that the next hypothesis extension must start with position b_k that is the next left of the last covered position in $\bar{c}$. The constraints popular in related work are IBM constraints (Berger et al., 1996). These allow only a limited number m of word positions to be skipped. If the coverage vector $\bar{c}$ already has m untranslated positions to the left of the last covered position j', then either the position $j'+1$ or one of these untranslated positions have to be translated next.

A constraint that we use in our system limits the number of *runs*. We define a *run* to be a contiguous sequence of covered positions (a sequence of 1's in the coverage vector). At any point in the search, we allow at maximum m runs. If m is set to 1, the search becomes monotonic. The run-based constraints are similar to the IBM constraints if they were to be defined in terms of the untranslated "gaps" of arbitrary length between the runs instead of the word positions.

3.3 Distortion penalty model

As one of the features in the loglinear translation model in Eq. 1 we include the reordering model.

The reordering model widely used in related research is a distance-based model. It assigns costs based on the distance from the end position of the last translated phrase to the start position of the current phrase; "jumps" over a long distance are penalized. The formula for the distortion model feature h_{Dist} is:

$$h_{Dist}(e_1^I, f_1^J) = \sum_{k=1}^{K+1} |j_{k-1} - b_k + 1| \qquad (2)$$

The sum in Equation 2 includes a jump from the beginning of the source sentence denoted by j_0 to the start of the first translated phrase, as well as the jump from the phrase translated last to the position b_{K+1} which we define as the position "one after the sentence end". It is clear from Equation 2 that the penalty for reordering increases linearly with increasing jump width. The influence of the model is controlled by the scaling factor λ_{Dist}.

We argue that the distance-based model is not the best choice even for the basic reordering model in statistical MT. The disadvantages of the model can be summarized as follows:

- The absolute distance is usually not a good indicator if a reordering should take place or not. From human translation experience, we know that long-range reorderings are less frequent than reorderings across a few words. However, it is not possible to say that skipping, e. g. 5 words for later translation is more probable than skipping 6 words, which is what the distance-based model suggests. A good example here is a typical Farsi sentence with a subject-object-verb (SOV) word ordering. For translation into English, the verb has to be translated directly after the subject so that the object has to be skipped for later translation. However, the object can be complex, and its length is more or less arbitrary.

- A hypothesis is penalized by the distance-based model not only for deviating from the monotonic translation path, but also for returning to it. Consider the following example coverage vector:

$$\mathbf{1\ddot{0}1111\dot{1}000000} \qquad (3)$$

The last translated position is marked here with a single dot, the first position of the next candidate source range is marked with two dots. We see here that by translating position 2 the system would fill the "gap" and would then continue to translate monotonically from position 8. Yet this intention is penalized by high costs of the jump from position 7 to position 2, and then back to position 8. This is clearly not the desired behavior, because the hypothesis in the example 3 competes with other hypotheses which do not fill the gap or even create new short gaps. All of these other hypotheses are not penalized by the distance-based model as harshly.

- The distance-based model penalizes linguistically very improbable reorderings less than the other, more reasonable reorderings. Consider the

following two permutations of a 9-word sentence, corresponding to the order in which the sentence positions are processed by the SCSS algorithm: 213546879 and 567812349. The first permutation includes many local reorderings which are highly unlikely. According to the second permutation, a block of 4 words starting from position 5 is translated first, then the first 4 words are translated, and the translation is continued with the 9th position. Such reordering may be necessary in language translation. Yet the total "jump" distance according to Eq. 2 is 12 for the first permutation, but 16 for the second one. Thus, when the scaling factor for the distance-based model is low, reorderings which are undesired according to human judgment and automatic evaluation measures like BLEU can be selected. That is why the minimum error rate training often assigns a high scaling factor to the model, so that both "good" and "bad" reorderings get a high penalty. This leads to a frequent phenomenon: automatic optimization of the loglinear model scaling factors leads to monotonic translation of almost every sentence.

3.4 Run-based penalty model

As a remedy to the problems described above, we introduce a novel model for reordering, which we call the *run-based* penalty model. This model introduces a penalty for each new run. The run concept was defined in Section 3.2. A penalty λ_{nr} is added only if the coverage vector of the new hypothesis has a higher number of runs than the coverage vector of the previous hypothesis. We define this to be the case when the position $b_k - 1$ to the left of the candidate start position b_k has not yet been covered. In such cases, a "new" non-monotonicity is introduced. To further distinguish between the runs, in practice we use three penalties: one for local reordering, when the new run is started from a position b_k that is no more than 3 positions away from the last covered position j_{k-1}; one for medium-range reordering when the starting position b_k of the new run is between 4 and 7 positions from j_{k-1}; and one for long-range reordering, used for jumps of more than 7 positions which start a new run. An optimal set of values for the 3 penalties can favor or discourage some of these reordering types.

In addition, we would like to penalize any deviation from the monotonic path in dependency on the length of the sentence part that has been translated non-monotonically in the current hypothesis. The question is, however, how to define such length. Defining it in terms of the number of skipped positions is not the best solution. An example that illustrates this is again a verb-final Farsi sentence: any good translator (MT or human) would first translate the verb, thus skipping many consecutive source positions. This should not be improbable, since in the next steps these positions will be translated monotonically. A better solution for measuring the length of the non-monotonic region is to determine the range in the coverage vector starting with the most left 0 and ending with the most right 1. Then, the "length" of the non-monotonicity can be defined as the minimum between the number of 0's and the number of 1's in this range. More formally, we define the range of source positions $(j_l, \ldots, j_r)$ with:

$$\begin{aligned} j_l &:= \min\{j : 1 \leq j \leq J \wedge c_j = 0\} \\ j_r &:= \max\{j : 1 \leq j \leq J \wedge c_j = 1\} \end{aligned} \quad (4)$$

The degree of deviation from monotonic translation path is then defined as:

$$r := \min\{|\{c_j = 0\}|, |\{c_j = 1\}| : j_l \leq j \leq j_r\} \quad (5)$$

The value r from Eq. 5 is used as another feature in the loglinear translation model with the scaling factor λ_r. Note that for both of the typical reorderings involving one word

$$\begin{aligned} &\mathbf{10\dot{1}\ddot{0}0000} \\ &\mathbf{1\ddot{0}0000\dot{1}0} \end{aligned} \quad (6)$$

the value r is equal to 1, so that the penalty introduced by this feature will be small. The second line in the example 6 corresponds to the example in Section 3.3, when the Farsi verb at the end of the sentence has to be translated after translating the subject.

The penalty r becomes large for e. g. the following coverage vector:

$$\mathbf{10101\dot{1}01\ddot{0}0000}$$

Here, there are many local reorderings, which, as already mentioned, is highly unlikely for translation between natural languages. The penalty will also become large if both the number of consecutive skipped positions and the length of the region that is translated monotonically after the skip is high:

$$\mathbf{1100000111\dot{1}\ddot{0}0000}$$

A high penalty for such hypotheses is also reasonable for most language pairs. Usually, we deal with SOV or VSO type sentences which have to be reordered to match the SVO target translation. Thus, only the verb group (usually not more than 2 words) is reordered. The swapping of the longer subject and object constituents that could correspond to the example above is not probable and therefore should get a high penalty.

3.5 Soft rule-based constraints

Another goal of our research is to incorporate rule-based reorderings of the source sentences into the SMT. Instead of preprocessing the source sentences using POS-based or parse-based hand-crafted reordering rules (described in Section 4), and then translating monotonically, we include the rules as "soft" reordering constraints in the SCSS algorithm. To this end, we save the permutation of each sentence as defined by the rules, and use it to introduce bonuses or penalties for the SCSS reorderings which respect or violate this permutation. More formally, let π represent the rule-based permutation of a source sentence:

$$\pi \quad : \quad j \rightarrow \pi(j) \in \{1, \ldots, J\},\ 1 \leq j \leq J$$

In the search, the pair (j', j) of the last covered position j' and the candidate start position j is considered "good" and assigned a bonus λ_g if the following equation holds:

$$\pi(j) = \pi(j') + 1 \tag{7}$$

Thus, the bonus is assigned if the two positions are translated in the same order as defined by the rule-based permutation. The same pair of positions is considered "bad" and assigned a penalty λ_b if Eq. 7 is not fulfilled, and in addition $j \neq j' + 1$. This means that no penalty is assigned in case of a monotonic translation of the two positions. Since the values λ_g and λ_b are included as features in the log-linear translation model for each starting position b_k in the search, the reordering path that is most favored by this model is the one that exactly corresponds to the permutation π. The path that includes reorderings which are not part of the rule-based permutation may be heavily penalized.

3.6 Reordering in training

Reordering of source sentences with hand-crafted rules may not always be reliable or useful for translation. That is why applying the rules to all training source sentences before word alignment and phrase extraction may harm translation quality. Instead, we decided to take only the source sentences which had actually been reordered after application of the rules, together with their target language counterparts, and add these data to the original training corpus. Thus, the word alignment learned with the iterative EM-based algorithm may be somewhat improved since it has to align both the re-ordered and the original version of a source sentence to the same target sentence. Experimental results for Farsi-to-Arabic in Section 5 show that including the reordered sentences is better in terms of automatic MT measures than using the original training corpus only.

4 Reordering Farsi sentences for MT

The Farsi language is considered to be a SOV language. The idea in reordering Farsi sentences is to mimic the word order in the target language. There are many reasons for the word order differences but the position of the verb is the most important one. In this study, we only propose rules for reordering verbs.

We use two approaches for the rule-based reordering of Farsi sentences before the SMT training, one based on POS tags and one based on parse trees. POS-based reordering rules utilize the Farsi POS tags and try to reorder the sentence according to the target language word order. POS-based reordering rules try to arrange words one-at-a-time in a way that fits best to the target sentence word order. Moving words one by one is dangerous as it can break the integrity of compound verbs which occur widely in Farsi. Therefore, an initial step of preprocessing is carried out to mark compound verbs in the sentence. This way, the entire compound verb is moved without breaking the integrity of the verb.

Another issue that needs special attention is the integrity of the clauses. A word in a clause should remain in it after the reordering. In order to achieve this, no word is moved beyond the clause borders which are designated with punctuation marks and conjunctions.

Parse tree based reordering rules try to reorder the Farsi sentences by moving phrases in a way that fits best to the target sentence phrase order. The reordering rules process the input parse tree in bracket notation and output the reordered sentence.

4.1 Farsi POS tagging

The POS tagger used in this study employs a Hidden Markov Model (HMM) approach. The Bayes decision rule for the tagger is formulated in the following well-known equation:

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-1}) \qquad (8)$$

In Eq. 8, the sequence of the words in the observed sentence are represented as $w_1 \dots w_n$. In order to calculate the estimate of the tag sequence $\hat{t}_1^n$, tag transition probabilities and word likelihoods are used. The tag transition probability $p(t_i|t_{i-1})$ is estimated with a bigram language model (LM) that is constructed from a tag transition corpus. The word likelihood $p(w_i|t_i)$ is computed from counts in the training data. Viterbi algorithm is used to determine the most likely tag sequence.

The Bijan-Khan corpus (Oroumchian et al., 2006) is used for building the models. The corpus contains about 2.6 million manually tagged words with a tag set containing 40 POS labels. The test file, which is 1% of the corpus, contains 894 sentences and close to 26K words of which 6,406 are unique. The out-of-vocabulary (OOV) rate in the test file is 1.5%. The system achieves an accuracy rate of 95.54%.

4.2 Farsi syntactic parsing

We used a unification-based active chart parser. The grammar rules are manually crafted in the *Lexical Functional Grammar* (LFG) paradigm. Unification is used as the fundamental mechanism to integrate information from lexical entries into larger grammatical constituents. The Farsi grammar contains 20 morphology rules and 45 syntax rules. The grammar also contains a lexicon of 67K entries. Each entry in the lexicon contains feature value pairs compulsory for morphological and syntactic analysis. Verbs and nouns in the lexicon contain appropriate inflection patterns besides other linguistic data.

4.3 Reordering rules for Farsi-to-Arabic translation

The main difference between Farsi word order and Arabic word order takes root from the place of the verb. VSO is a widely used word order in Arabic and the verb is generally in a sentence initial position, whereas in Farsi it is followed by the object in a sentence final position. Therefore, the main aim in the Farsi-to-Arabic reordering rules is to move the Farsi verb to a sentence initial position.

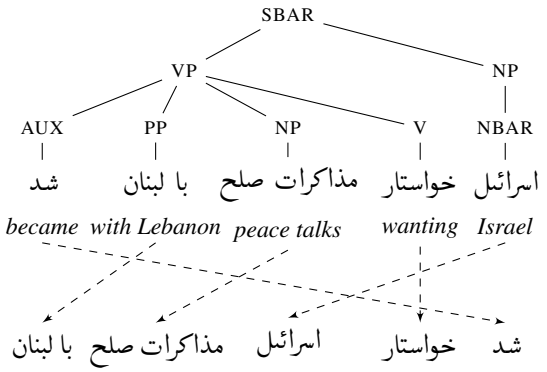

Figure 1: Parse-based reordering of the Farsi sentence in (1).

An example POS-based reordering is presented for the Farsi sentence with the reference translation "Israel calls for peace talks with Lebanon" in (1). The reordered form of the same sentence is given in (2) in which the verb is moved to the beginning of the sentence. The corresponding Arabic sentence is given in (3).

(1) شد لبنان با صلح مذاکرات خواستار اسرائیل
AUX N-SG PREP N-SG N-PL V-ING N-SG
became Lebanon with peace talks wanting Israel

(2) لبنان با صلح مذاکرات اسرائیل خواستار شد
N-SG PREP N-SG N-PL N-SG V-ING AUX
Lebanon with peace talks Israel wanting became

(3) لبنان مع سلام لمحادثات اسرائیل تدعو
Lebanon with peace for talks Israel calls

Parse-based reordering for the sentence in (1) is depicted in Figure 1.

4.4 Reordering rules for Farsi-to-English translation

Reordering rules for Farsi-to-English are implemented in a similar manner as in Farsi-to-Arabic translation. This time the verb is moved to a position directly after the subject in the phrase since the sentences in English conform to the SVO word order. The reordered form of the sentence in (1) is given in (4).

Table 1: Corpus statistics of the training and test data.

	Farsi-Arabic	Farsi-English
Train: Sentences	289K	130K
Running Words	5.5M/4.4M	2.8M/2.6M
Vocabulary	94K/153K	60K/54K
+ affix splitting	94K/87K	–
Test: Sentences	536	536
Running Words	16K	16K
OOV rate (%)	0.9%	1.0%

(4) لبنان با صلح مذاکرات خواستار شد اسرائیل
N-SG PREP N-SG N-PL V AUX N-SG
Lebanon with peace talks wanting became Israel
'Israel calls for peace talks with Lebanon'

We only used parse-based rules for Farsi-to-English translation because the task of finding the last word of the subject based only on POS tags is too ambiguous.

5 Experiments

5.1 Experimental setup

The experimental evaluation of the proposed methods was performed on a Farsi-to-English and a Farsi-to-Arabic translation task, using in-house data for training. The corpus statistics of the training and test data for the two tasks are summarized in Table 1.

The Farsi side of the training and test data was compiled from sources which include articles from newspapers, news websites, commercial companies and public organizations, etc.. These data was then translated into English and Arabic by linguists in-house and by translation agencies.

The preprocessing for all 3 languages included tokenization and categorization of numbers. For Arabic, we additionally used morphological segmentation similar to the ATB-style segmentation (Maamouri et al., 2004). This segmentation helped to reduce the Arabic vocabulary size from 153K to 87K and make it closer in size to the Farsi vocabulary. The segmentation was removed after translation by concatenating the detached prefixes and suffixes marked with "_" with the following or preceding word, respectively.

The phrase table was extracted using GIZA++ word alignments (Och and Ney, 2003) and the phrase extraction implemented in the Moses statistical MT toolkit (Koehn et al., 2007). In translation to Arabic, we used two language models: a trigram LM trained on the target part of the bilingual data, and a 5-gram LM trained on the Arabic Gigaword corpus (720M running words). Both Arabic LMs were estimated using the data with morphological segmentation as described above. For English, we used a huge 5-gram LM trained on the English Gigaword corpus and additional in-house data (3.9 billion words). The LMs were trained using the IRSTLM toolkit (Federico et al., 2008).

Table 2: Farsi-to-English translation results using different reordering models.

Reordering model	BLEU (%)	WER (%)
monotonic	29.1	66.3
jump-based penalty	29.4	66.0
run-based penalty	30.6	65.6
parse-based hard	28.9	66.4
parse-based soft	30.5	64.9

The 536-sentence test corpus described in Table 1 was used for both translation tasks. Three reference translations were created independently for each sentence in this corpus. We divided the corpus at random into two 268-sentence-long parts, with the goal of using one of the parts as the development set for optimization of the log-linear model scaling factors. In order to avoid over-fitting, we did separate optimizations on each of the two halves, and then extracted the hypotheses for the other half using the optimized parameters. The results reported in this section are obtained by automatically scoring the concatenation of the two sets of translation hypotheses obtained in this way.

The evaluation was performed using the well-established automatic measure BLEU (Papineni et al., 2002). In addition, we computed the multiple-reference word error rate (WER). WER penalizes wrong word order in translations. That is why it is a good choice to rate any improvement caused by new reordering models. The evaluation was case-insensitive. The objective function for the parameter optimization described above was chosen to be the average between $1-$BLEU and WER.

5.2 Farsi-to-English MT

The experimental results for the Farsi-to-English MT system are shown in Table 2. There are two

Table 3: Examples of improved translation quality using the run-based penalty vs. the jump-based penalty model. Reference translations are marked with REF.

REF:	Ahmadinejad: "Bullying" powers cannot stop Iran's nuclear program.
JMP:	Ahmadinejad said powers "bullying" can not nuclear program to stop Iran.
RUN:	Ahmadinejad said powers "bullying" Iran's nuclear program can not stop.
REF:	A few days ago I called upon their leader...
JMP:	I was a few days ago to the leader...
RUN:	A few days ago I told their leader...
REF:	We (the US) have permanent friends, but we don't have permanent enemies.
JMP:	Our US friends forever, but permanent enemies have.
RUN:	We the United States friends forever, but we do not have permanent enemies.

baselines to which we compare our method: the monotonic translation and the translation using the well-established distance-based penalty model. We see that the positive influence of the distance-based model is small: the BLEU score improves from 29.1 to 29.4%, the same absolute improvement is achieved in terms of WER. When we now use the run-based penalty model with the 4 features described in Section 3.4, we improve the BLEU score by 1.5% absolute and WER by 0.7% absolute as compared to the monotonic translation. Thus, the run-based penalty model clearly outperforms the distance-based penalty model. Checking the feature scaling factors for the short, medium, and long range new run penalties after the optimization, we observed that the factor for the short-range feature was small, but negative, thus assigning a bonus to local reorderings. In contrast, the penalty for a long-range jump was 10 times higher than for a medium-range one. Examples of improved translanslation word order and quality when using the run-based penalty model instead of the distance-based model are presented in Table 3.

Next, we tested the application of parse-based reordering rules. Applying these rules to the source sentences and then performing monotonic SMT (the "hard" reordering in Table 2) resulted in a degradation of the MT error measures. We attribute this to the ambiguity of the verb reordering task: the goal was to put the verb between subject and object, but the detection of the boundary between subject and object is a hard task at which the parser often seems to fail.

Table 4: Effect of adding reordered Farsi sentences and their Arabic counterparts to the training data (reordering in the search performed using run-based penalties).

System	BLEU (%)	WER (%)
baseline	19.9	70.6
+ reordering in train.	21.1	70.0

Including the reorderings proposed by the rules as the "soft" penalty-type constraints in the SMT search (see Section 3.5) leads to better results. When combined with the run-based penalty model, the additional features for "good" and "bad" parse-based reorderings help to achieve a further reduction of WER from 65.6 to 64.9% as compared to using the run-based model only, while the BLEU score remained almost unchanged. Thus, the total reduction in WER as compared to the monotonic baseline is 1.4% absolute.

5.3 Farsi-to-Arabic MT

For the Farsi-to-Arabic translation task, we first show the effect of adding reordered Farsi training sentences with their target language counterparts to the original bilingual training corpus as described in Section 3.6. Table 4 compares the baseline system trained on the unreordered corpus with this system. About 52% of the sentences were actually reordered with parse-based rules; we did not use POS-based rules for reordering. Thus, the training corpus size increased from 289K to 439K. In translation, we used the run-based penalty model for this experiment. The improvement in MT error measures over the baseline is substantial: 1.2% absolute in BLEU and 0.6% absolute in WER. Interestingly, we did not observe any notable improvement when performing the same experiment for Farsi-to-English translation. We speculate that the main reason for the improvement here is better word alignment quality: because of its sentence-final position, the Farsi verb is often not aligned to its Arabic counterpart in the sentence-initial position due to alignment model restrictions[1]. Having seen the same verb in two positions, the iterative alignment algorithm may be able to align it correctly.

It is interesting to observe that on the same test

[1]such as the distortion model interpolated with a distance-based penalty

Table 5: Farsi-to-Arabic translation results using different reordering models (using reordering in training as in Table 4).

Reordering model	BLEU (%)	WER (%)
monotonic	20.7	70.2
jump-based penalty	20.4	71.4
run-based penalty	21.1	70.0
POS-based hard	18.3	74.4
POS-based soft	20.4	70.0
parse-based hard	20.1	70.8
parse-based soft	20.2	70.0

set, the translation quality for Farsi-to-English is by 10% absolute better than for Farsi-to-Arabic, although we have about twice as much training data available for the Farsi-to-Arabic task. We attribute this to the structural closeness between Farsi and English in comparison to the structural difference between Farsi and Arabic. Better quality of the Farsi-to-English human translations of the training and test data might be another reason. Also, the training data for Farsi-to-Arabic is less homogeneous, which results in a large vocabulary size of 94K words.

The experiments with different reordering models for Farsi-to-Arabic are summarized in Table 5. Here, the distance-based penalty model is not able to improve the translation quality as compared to the monotonic translation. The scaling factor for the model is optimized to be quite high, so that only a few sentences are reordered at all. In contrast, the novel run-based penalty model is able to improve the monotonic baseline slightly by 0.4% absolute in BLEU and 0.2% absolute in WER. Interestingly, the model favored medium-range reorderings, whereas short-range new runs were penalized 8 times stronger than the long-range ones.

Next, we experimented with both POS-based and parse-based reordering rules. Unfortunately, neither the "hard" or the "soft" way of introducing rule-based reorderings into the SMT search leads to improvement of the MT quality on this task. The reasons for this have to be further investigated. We suspect that the POS-based reordering rules scramble the sentence in an irrecoverable manner. The appropriateness of the POS tag set is another issue. The compound verbs in Farsi introduce further complications to the reordering process. However, parse-based reorderings perform similar to the baseline monotonic model when applied either in a preprocessing step or used to define reordering penalties in the search. This shows that the parse-based rules, if further enhanced, have the potential of improving MT quality. Another conclusion we can make from Table 5 is that even if the rules suggest incorrect reorderings (this is often the case for the POS-based rules), the "soft" way of introducing these rules helps the MT system to recover from these errors. This is achieved in the process of optimizing the model scaling factors on the development data. In fact, the factor for the rule-based reorderings was optimized not to be negative (which would have meant a bonus for them), but positive. However, the penalty factor for the other "bad" reorderings not compliant with the rules (as described in Section 3.6) was determined to be about 5 times larger.

6 Conclusions

In this paper, we proposed a novel model for scoring reordering in phrase-based SMT. Instead of using the absolute distance between the last covered and the current source position to define the reordering penalty, we introduced penalties for starting a new translation run (a new deviation from a monotonic translation path), as well as for increasing the total number of source positions in a partial MT hypothesis which are translated non-monotonically. In addition, we successfully incorporated hand-crafted reordering rules into the SMT search by introducing penalties for those partial sentence permutations in the SMT search which are not part of the permutation defined by the rules. We applied the proposed methods to translation tasks from Farsi into English and Arabic. We described in detail the Farsi reordering rules which are based on either POS tags or syntactic parses.

In the experiments on general-domain data with large vocabularies, we could show that the run-based penalty model results in significant improvement in terms of automatic MT error measures on the Farsi-to-English task as compared to the monotonic translation baseline. On both Farsi-to-Arabic and Farsi-to-English tasks, the model outperforms the distance-based penalty model. Introducing parse-based reordering rules as hard constraints did not help to further improve MT quality, which we attribute to the insufficient quality of the rules. How-

ever, using the rules as soft constraints in the SMT search together with the run-based penalty model may have alleviated the negative effects of the rules, whereas their positive effects were in some cases taken into account.

In the future, we would like to test the proposed system with more effective reordering rules. In addition, we plan to extend the run-based penalty model by introducing the dependency on the source words which trigger a new run.

References

Y. Al-Onaizan and K. Papineni. 2006. Distortion models for statistical machine translation. In *Proc. of ACL*, pages 529–536.

J.W. Amtrup, K. Megerdoomian, and R. Zaja. 2000. Rapid Development of Translation Tools: Application to Persian and Turkish. In *Proc. of COLING*, volume 2, pages 982 – 986, Saarbrücken, Germany.

A.L. Berger, P.F. Brown, S.A.D. Pietra, V.J.D. Pietra, J.R. Gillett, A.S. Kehler, and R.L. Mercer. 1996. Language translation apparatus and method of using context-based translation models, United States Patent 5510981, April.

B. Chen, M. Cettolo, and M. Federico. 2006. Reordering rules for phrase-based statistical machine translation. In *Proc. of IWSLT*, pages 1–15.

M. Collins, P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proc. of ACL*, pages 531–540.

Y. Deng and B. Zhou. 2009. Optimizing word alignment combination for phrase table training. In *Proc. of the ACL-IJCNLP 2009*, pages 229–232, Suntec, Singapore, August.

M. Federico, N. Bertoldi, and M. Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech*, Brisbane, Australia, September.

C.L. Kao, S. Saleem, R. Prasad, F. Choi, P. Natarajan, D. Stallard, K. Krstovski, and M. Kamali. 2008. Rapid Development of an English/Farsi Speech-to-Speech Translation System. In *Proc. of IWSLT*, pages 166–173, Hawaii, USA.

A. Kathol and J. Zheng. 2008. Strategies for building a Farsi-English SMT system from limited resources. In *Interspeech*, pages 2731–2734.

P. Koehn, A. Axelrod, A.B. Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *Proc. of IWSLT*, October.

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 *Proc. of ACL*, Prague, Czech Republic.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *AMTA 04*, pages 115–124, Washington DC, September/October.

C.H. Li, D. Zhang, M. Li, M. Zhou, M. Li, and Y. Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proc. of ACL*, pages 720–727, June.

M. Maamouri, A. Bies, T. Buckwalter, and W. Mekki. 2004. The Penn Arabic treebank: Building a large-scale annotated Arabic corpus. In *Proc. of the NEMLAR Conference on Arabic Language Resources and Tools*, pages 102–109, Cairo, Egypt.

M. Nagata, K. Saito, K. Yamamoto, and K. Ohashi. 2006. A clustered global phrase reordering model for statistical machine translation. In *Proc. of ACL*, page 720.

F.J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

F.J. Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167, Sapporo, Japan, July.

F. Oroumchian, S. Tasharofi, H. Amiri, H. Hojjat, and F. Raja. 2006. Creating a feasible corpus for Persian POS tagging. Technical Report TR3/06, University of Wollongong in Dubai.

K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of ACL*, pages 311–318, Morristown, NJ, USA.

K. Rottmann and S. Vogel. 2007. Word reordering in statistical machine translation with a POS-based distortion model. In *Proc. of TMI*.

C. Saedi, Y. Motazadi, and M. Shamsfard. 2009. Automatic translation between English and Persian texts. In *CAASL3: Proc. of the 3rd Workshop on Computational Approaches to Arabic-script based Languages*, Ottawa, Ontario, Canada.

C. Wang, M. Collins, and P. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *In Proc. of EMNLP*, pages 737–745.

R. Zens and H. Ney. 2006. Discriminative Reordering Models for Statistical Machine Translation. In *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 55–63, New York City, NY, June.

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

Y. Zhang, R. Zens, and H. Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *Proc. of IWSLT*, pages 21–28, Trento, Italy, October.