Syntactic Reordering for English-Arabic Phrase-Based Machine Translation

Jakob Elming
Languagelens
Copenhagen, Denmark
je@languagelens.com

Nizar Habash
Center for Computational Learning Systems
Columbia University, New York, USA
habash@ccls.columbia.edu

Abstract

We investigate syntactic reordering within an English to Arabic translation task. We extend a pre-translation syntactic reordering approach developed on a close language pair (English-Danish) to the distant language pair, English-Arabic. We achieve significant improvements in translation quality over related approaches, measured by manual as well as automatic evaluations. These results prove the viability of this approach for distant languages.

1 Introduction

The emergence of phrase-based statistical machine translation (PSMT) (Koehn et al., 2003a) has been one of the major developments in statistical approaches to translation. Allowing translation of word sequences (phrases) instead of single words provides PSMT with a high degree of robustness in word selection and in local-word reordering. Recent developments have shown that improvements in PSMT quality are possible using syntax. One such development is the pre-translation reordering approach, which adjusts the source sentence to resemble target-language word order prior to translation. This is typically done using rules that are either manually created or automatically learned from word-aligned parallel corpora.

One particular variety of this approach, proposed by Elming (2008), uses a large set of linguistic features to automatically learn reordering rules. The rules are applied non-deterministically; however, phrase-internal word-alignments are used to ensure that the intended reordering does not come undone because of phrase internal reordering (Elming, 2008). This approach was shown to produce improved MT output on English-Danish MT, a relatively closely-related and similarly-structured language pair. In this paper, we study whether this approach can be extended to distant language pairs, specifically English-to-Arabic. We achieve significant improvement in translation quality over related approaches, measured by manual as well as automatic evaluations on this task. This proves the viability of this approach on distant languages. We also examined the effect of the alignment method on learning reordering rules. Interestingly, our experiments produced better translation using rules learned from automatic alignments than using rules learned from manual alignments.

In the next section, we discuss and contrast related work. Section 3 describes aspects of English and Arabic structure that are relevant to reordering. Section 4 describes the automatic induction of reordering rules and its integration in PSMT. In section 5, we describe the SMT system used in the experiments. In section 6, we evaluate and discuss the results of our English-Arabic MT system.

2 Related Work

Much work has been done in syntactic reordering for SMT, focusing on both source and target-language syntax. In this paper, we adapt an approach that utilizes source-syntax information as opposed to target-side syntax systems (Yamada and Knight, 2001; Galley et al., 2004). This is because we are translating from English to Arabic and we are discouraged by recent results indicating Arabic parsing is not at a stage that makes it usable in MT (Habash et al., 2006). While several recent authors using a pre-translation (source-side) reordering approach have achieved positive results, it has been difficult to integrate syntactic

Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages, pages 69–77,
Athens, Greece, 31 March, 2009.

information while retaining the strengths of the statistical approach. In some studies, reordering decisions are done "deterministically" by supplying the decoder with a canonical word order (Xia and McCord, 2004; Collins et al., 2005; Wang et al., 2007; Habash, 2007). These reordering rules are either manually specified or automatically learned from alignments; and they are always placed outside the actual PSMT system. By contrast, other studies (Crego and Mariño, 2007; Zhang et al., 2007; Li et al., 2007; Elming, 2008) are more in the spirit of PSMT, in that multiple reorderings are presented to the PSMT system as (possibly weighted) options that are allowed to contribute alongside other parameters. Specifically, we follow the pre-translation reordering approach of Elming (2008). This approach has been proven to remedy shortcomings of other pre-translation reordering approaches by reordering the input word sequence, but scoring the output word sequence.

Elming (2008) only examined the approach within English – Danish, a language pair that displays little reordering. By contrast, in this paper, we target the more demanding reordering task of translating between two distant languages, English and Arabic. While much work has been done on Arabic to English MT (Habash and Sadat, 2006; Lee, 2004) mostly focusing on addressing the problems caused by the rich morphology of Arabic, we handle the less described translation direction: English to Arabic. Recently, there are some new publications on English to Arabic MT. Sarikaya and Deng (2007) use joint morphological-lexical language models to re-rank the output of English dialectal-Arabic MT, and Badr et al. (2008) report results on the value of the morphological decomposition of Arabic during training and describe different techniques for re-composition of Arabic in the output. We differ from the previous efforts targeting Arabic in that (1) we do not address morphology issues through segmentation (more on this in section 3) and (2) we focus on utilizing syntactic knowledge to address the reordering challenges of this translation direction.

3 Arabic Syntactic Issues

Arabic is a morphologically and syntactically complex language with many differences from English. Arabic morphology has been well studied in the context of MT. Previous results all suggest that some degree of tokenization is helpful when translating from Arabic (Habash and Sadat, 2006; Lee, 2004). However, when translating into a morphologically rich language, target tokenization means that the translation process is broken into multiple steps (Badr et al., 2008). For our experiments, Arabic was not segmented apart from simple punctuation tokenization. This low level of segmentation was maintained in order to agree with the segmentation provided in the manually aligned corpus we used to learn our rules (section 6.1). We found no simple means for transferring the manual alignments to more segmented language. We expect that better performance would be achieved by introducing more Arabic segmentation as reported by Badr et al. (2008).[1] As such, and unlike previous work in PSMT translating into Arabic, we focus here on syntax. We plan to investigate different tokenization schemes for syntactic preprocessing in future work. Next, we describe three prominent English-Arabic syntactic phenomena that have motivated some of our decisions in this paper.

First is verb-subject order. Arabic verb subjects may be: (a.) pro-dropped (verb conjugated), (b.) pre-verbal (SVO), or (c.) post-verbal (VSO). Although the English SVO order is possible in Arabic, it is not always preferred, especially when the subject is particularly long. Unfortunately, this is the harder case for PSMT to handle. For small subject noun phrases (NP), PSMT might be able to handle the reordering in the phrase table if the verb and subject were seen in training. But this becomes much less likely with very long NPs that exceed the size of the phrases in a phrase table. The example in Figure 1 illustrates this point. Bolding and italics are used to mark the verb and subordinating conjunction that surround the subject NP (19 tokens) in English and what they map to in Arabic, respectively.[2]

Secondly, Arabic adjectival modifiers typically follow their nouns with the exception of some superlative adjectives. However, English adjectival modifiers can follow or precede their nouns depending on the size of the adjectival phrase: single word adjectives precede but multi-word adjectives phrases follow (or precede while hyphenated). For example, *a tall man* translates as رجل طويل *rjl*

[1] Our results are not comparable to their results, since they report on non-standard data sets.

[2] All Arabic transliterations in this paper are provided in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

[NP-SBJ The general coordinator of the railroad project among the countries of the Gulf Cooperation Council , Hamid Khaja ,] **[V announced]** *[SUB that ...]*

[V اعلن] [NP-SBJ المنسق العام لمشروع السكة الحديد بين دول مجلس التعاون الخليجي حامد خاجه] [SUB ان ...]

[V Aʕln] [NP-SBJ Almnsq AlʕAm lmšrwʕ Alskħ AlHdyd byn dwl mjls AltʕAwn Alxlyjy HAmd xAjh] *[SUB An ...]*

Figure 1: An example of long distance reordering of English SVO order to Arabic VSO order

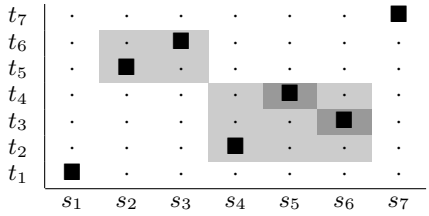

Figure 2: Abstract alignment matrix example of reordering.

Twyl 'man tall'; however, the English phrase *a man tall of stature* translates with no reordering as رجل طويل القامة *rjl Twyl AlqAmħ* 'man tall the-stature'. So does the superlative *the tallest man* translating into اطول رجل *ATwl rjl* 'tallest man.'

Finally, Arabic has one syntactic construction, called *Idafa*, for indicating possession and compounding, while English has three. The Idafa construction typically consists of one or more indefinite nouns followed by a definite noun. For example, the English phrases *the car keys*, *the car's keys* and *the keys of the car* all translate into the Arabic مفاتيح السيارة *mfAtyH AlsyArħ* 'keys the-car.' Only one of the three English constructions does not require content word reordering.

4 Reordering rules

4.1 Definition of reordering

Following Elming (2008), we define reordering as two word sequences, left sequence (LS) and right sequence (RS), exchanging positions. These two sequences are restricted by being parallel consecutive, maximal and adjacent. The sequences are not restricted in length, making both short and long distance reordering possible. Furthermore, they need not be phrases in the sense that they appear as an entry in the phrase table.

Figure 2 illustrates reordering in a word alignment matrix. The matrix contains reorderings between the light grey sequences (s_2^3 and s_4^6)[3] and the dark grey sequences (s_5^5 and s_6^6). On the other hand, the sequences s_3^3 and s_4^5 are not considered for reordering, since neither one is maximal, and s_4^5 is not consecutive on the target side.

4.2 Learning rules

Table 1 contains an example of the features available to the algorithm learning reordering rules. We include features for the candidate reordering sequences (LS and RS) and for their possible left (LC) and right (RC) contexts. In addition to words and parts-of-speech (POS), we provide phrase structure (PS) sequences and subordination information (SUBORD). The PS sequence is made up of the highest level nodes in the syntax tree that cover the words of the current sequence and only these. Subordinate information can also be extracted from the syntax tree. A subordinate clause is defined as inside an SBAR constituent; otherwise it is a main clause. Our intuition is that all these features will allow us to learn the best rules possible to address the phenomena discussed in section 3 at the right level of generality.

In order to minimize the amount of training data, word and POS sequences are annotated as too long (T/L) if they are longer than 4 words, and the same for phrase structure (PS) sequences if they are longer than 3 units. A feature vector is only used if at least one of these three levels is not T/L for both LS and RS, and T/L contexts are not included in the set. This does not constrain the possible length of a reordering, since a PS sequence of length 1 can cover an entire sentence. In the example in Table 1, LS and RS are single words, but they are not restricted in length. The span of the contexts varies from a single neighboring word to all the way to the sentence border. In the example, LS and RS should be reordered, since adjectives appear as post-modifiers in Arabic.

In order to learn rules from the annotated data, we use a rule-based classifier, Ripper (Cohen,

[3]Notation: s_x^y means the consecutive source sequence covering word positions x to y.

Level	LC	LS	RS	RC
WORD	<s> he bought || he bought || bought	new	books	today || today . || today . < /s>
POS	<S> NN VBD || NN VBD || VBD	JJ	NNS	NN || NN . || NN . < /S>
PS	<S> NP VBD || NP VBD || VBD	JJ	NNS	NP || NP . || NP . < /S>
SUBORD	MAIN	MAIN	MAIN	MAIN

Table 1: Example of features for rule-learning. Possible contexts separated by ||.

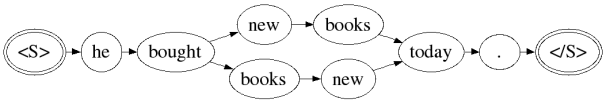

Figure 3: Example word lattice.

1996). The motivation for using Ripper is that it allows features to be sets of strings, which fits well with our representation of the context, and it produces easily readable rules that allow better understanding of the decisions being made. In section 6.3, extracted rules are exemplified and analyzed.

The probabilities of the rules are estimated using Maximum Likelihood Estimation based on the information supplied by Ripper on the performance of the individual rules on the training data. These logarithmic probabilities are easily integratable in the log-linear PSMT model as an additional parameter by simple addition.

5 The PSMT system

Our baseline is the PSMT system used for the 2006 NAACL SMT workshop (Koehn and Monz, 2006) with phrase length 3 and a trigram language model (Stolcke, 2002). The decoder used for the baseline system is Pharaoh (Koehn, 2004) with its distance-penalizing reordering model. Since Pharaoh does not support word lattice input, we use our own decoder for the experiments. Except for the reordering model, it uses the same knowledge sources as Pharaoh, i.e. a bidirectional phrase translation model, a lexical weight model, phrase and word penalties, and a target language model. Its behavior is comparable to Pharaoh when doing monotone decoding.

The search algorithm of our decoder is similar to the RG graph decoder of (Zens et al., 2002). It expects a word lattice as input. Figure 3 shows the word lattice for the example in table 2. In the example used here, we choose to focus on the reordering of adjective and noun. For readability, we do not describe the possibility of reordering the subject and verb. This will also be the case in later use of the example.

Since the input format defines all possible word orders allowed by the rule set, a simple monotone search is sufficient. Using a language model of order n, for each hypothesized target string ending in the same n-1-gram, we only have to extend the highest scoring hypothesis. None of the others can possibly outperform this one later on. This is because the maximal context evaluating a phrase extending this hypothesis, is the history (n-1-gram) of the first word of that phrase. The decoder is not able to look any further back at the preceding string.

5.1 The reordering approach

Similar to Elming (2008), the integration of the rule-based reordering in our PSMT system is carried out in two separate stages:

1. Reordering the source sentence to assimilate the word order of the target language.
2. Weighting of the target word order according to the rules.

Stage (1) is done in a non-deterministic fashion by generating a word lattice as input. This way, the system has both the original word order, and the reorderings predicted by the rule set. The different paths of the word lattice are merely given as equal suggestions to the decoder. They are in no way individually weighted.

Separating stage (2) from stage (1) is motivated by the fact that reordering can have two distinct origins. They can occur because of stage (1), i.e. the lattice reordering of the original English word order (phrase *external* reordering), and they can occur inside a single phrase (phrase *internal* reordering). The focus of this approach lies in doing *phrase-independent* word reordering. Rule-predicted reorderings should be promoted regardless of whether they owe their existence to a syntactic rule or a phrase table entry.

This is accomplished by letting the actual scoring of the reordering focus on the target string.

Source:	he$_1$ bought$_2$ new$_3$ books$_4$ today$_5$	
Rule:	3 4 → 4 3	
Hypothesis	Target string	Alignment
H1	Aštrý jdydħ ktbA	1+2 3 4
H2	Aštrý ktbA jdydħ	1+2 4 3

Table 2: Example of the scoring approach during decoding at source word 4.

The decoder is informed of where a rule has predicted a reordering, how much it costs to do the reordering, and how much it costs to avoid it. This is then checked for each hypothesized target string via a word alignment.

The word alignment keeps track of which source position the word in each target position originates from. In order to access this information, each phrase table entry is annotated with its internal word alignment, which is available as an intermediate product from phrase table creation. If a phrase pair has multiple word alignments, the most frequent one is chosen.

Table 2 exemplifies the scoring approach, again with focus on the adjective-noun reordering. The source sentence is *'he bought new books today'*, and a rule has predicted that source word 3 and 4 should change place. Due to the pro-drop nature of Arabic, the first Arabic word is linked to the two first English words (1+2). When the decoder has covered the first four input words, two of the hypothesis target strings might be H1 and H2. At this point, it becomes apparent that H2 contains the desired reordering (namely what corresponds to source word order '4 3'), and it get assigned the reordering cost. H1 does not contain the rule-suggested reordering (instead, the words are in the original order '3 4'), and it gets the violation cost. Both these scorings are performed in a phrase-independent manner. The decoder assigns the reordering cost to H2 without knowing whether the reordering is internal (due to a phrase table entry) or external (due to a syntactic rule).

Phrase internal reorderings at other points of the sentence, i.e. points that are not covered by a rule, are not judged by the reordering model. Our rule extraction does not learn every possible reordering between the two languages, but only the most general ones. If no rule has an opinion at a certain point in a sentence, the decoder is free to choose the phrase translation it prefers without reordering cost.

Separating the scoring from the source language reordering also has the advantage that the approach in essence is compatible with other approaches such as a traditional PSMT system (Koehn et al., 2003b) or a hierarchical phrase system (Chiang, 2005). We will, however, not examine this possibility further in the present paper.

6 Evaluation

6.1 Data

We learn the reordering rules from the IBM Arabic-English aligned corpus (IBMAC) (Ittycheriah and Roukos, 2005). Of its total 13.9K sentence pairs, we only use 8.8K sentences because the rest of the corpus uses different normalizations for numerals that make the two sets incompatible. 6.6K of the sentences (179K English and 146K Arabic words) are used to learn rule, while the rest are used for development purposes. In addition to the manual alignment supplied with these data, we create an automatic word alignment for them using GIZA++ (Och and Ney, 2003) and the *grow-diag-final* (GDF) symmetrization algorithm (Koehn et al., 2005). This was done together with the data used to train the MT system. The English side is parsed using a state-of-the-art statistical English parser (Charniak, 2000). Two rule sets are learned based on the manual alignments (MAN) and the automatic alignments (GDF).

The MT system is trained on a corpus consisting of 126K sentences with 4.2M English and 3.3M Arabic words in simple tokenization scheme. The domain is newswire (LDC-NEWS) taken from Arabic News (LDC2004T17), eTIRR (LDC2004E72), English translation of Arabic Treebank (LDC2005E46), and Ummah (LDC2004T18). Although there are additional corpora available, we restricted ourselves to this set to allow for a fast development cycle. We plan to extend the data size in the future. The Arabic language model is trained on the 5.4M sentences (133M words) of newswire text in the 1994 to 1996 part of the Arabic Gigaword corpus. We restricted ourselves to this part, since we are not able to run Pharaoh with a larger language model.[4]

For test data, we used NIST MTEval test sets from 2004 (MT04) and 2005 (MT05)[5]. Since these data sets are created for Arabic-English evaluation with four English reference sentences for

[4] All of the training data we use is available from the Linguistic Data Consortium (LDC): http://www.ldc.upenn.edu/.

[5] http://www.nist.gov/speech/tests/mt/

System		Dev	MT04	MT05
Pharaoh Free		28.37	23.53	24.79
Pharaoh DL4		29.52	24.72	25.88
Pharaoh Monotone		27.93	23.55	24.72
MAN	NO weight	29.53	24.72	25.82
	SO weight	29.43	24.74	25.82
	TO weight	29.40	24.78	25.93
GDF	NO weight	29.87	25.11	26.04
	SO weight	29.84	25.06	26.01
	TO weight	29.95	25.17	26.09

Table 3: Automatic evaluation scores for different systems using rules extracted from manual alignments (MAN) and automatic alignments (GDF). The TO system using GDF rules is significantly better than the light grey cells at a 95% confidence level (Zhang et al., 2004).

each Arabic sentence, we invert the sets by concatenating all English sentences to one file. This means that the Arabic reference file contains four duplicates of each sentence. Each duplicate is the reference of a different English source sentence. Following this merger, MT04 consists of 5.4K sentences with 193K English and 144K Arabic words, and MT05 consists of 4.2K sentences with 143K English and 114K Arabic words. MT04 is a mix of domains containing speeches, editorials and newswire texts. On the other hand, MT05 is only newswire.

The NIST MTEval test set from 2002 (MT02) is split into a tuning set for optimizing decoder parameter weights and a development set for ongoing experimentation. The same merging procedure as for MT04 and MT05 is employed. This results in a tune set of 1.0K sentences with 34K English and 26K Arabic words, and a development set of 3.1K sentences with 102K English and 79K Arabic words.

6.2 Results and discussion

The reordering approach is evaluated on the MT04 and MT05 test sets. Results are listed in table 3 along with results on the development set. We report on (a) Pharaoh with no restriction on reordering (Pharaoh Free), (b) Pharaoh with distortion limit 4 (Pharaoh DL4), (c) Pharaoh with monotone decoding (Pharaoh Monotone), and (d) a system provided with a rule reordered word lattice but no (NO) weighting in the spirit of (Crego and Mariño, 2007), (e) the same system but with a source order (SO) weighting in the spirit of (Zhang et al., 2007; Li et al., 2007), and finally (f) the same system but with the target order (TO) weighting.

System	MT04	MT05	Avr. human
Pharaoh Free	24.07	25.15	3.0 (2.80)
Pharaoh DL4	25.42	26.51	—
NO scoring	25.68	26.29	2.5 (2.43)
SO scoring	25.42	26.02	2.5 (2.64)
TO scoring	25.98	26.49	2.0 (2.08)

Table 4: Evaluation on the diff set. Average human ratings are medians with means in parenthesis, lower scores are better, 1 is the best score.

In addition to evaluating the reordering approaches, we also report on supplying them with different reordering rule sets: a set that was learned on manually aligned data (MAN), and a set learned on the same data but with automatic alignments (GDF).

6.2.1 Overall Results

Pharaoh Monotone performs similarly to Pharaoh Free. This shows that the question of improved reordering is not about quantity, but rather quality: what constraints are optimal to generate the best word order. The TO approach gets an increase over Pharaoh Free of 1.3 and 1.6 %BLEU on the test sets, and 0.2 and 0.5 %BLEU over Pharaoh DL4.

Improvement is less noticeable over the other pre-translation reordering approaches (NO and SO). A possible explanation is that the rules do not apply very often, in combination with the fact that the approaches often behave alike. The difference in SO and TO scoring only leads to a difference in translation in $\sim$14% of the sentences. This set, the *diff* set, is interesting, since it provides a focus on the difference between these approaches. In table 4, we evaluate on this set.

6.2.2 Diff Set

Overall the TO approach seems to be a superior reordering method. To back this observation, 50 sentences of MT04 are manually evaluated by a native speaker of Arabic. Callison-Burch et al. (2007) show that ranking sentences gives higher inter-annotator agreement than scoring adequacy and fluency. We therefore employ this evaluation method, asking the evaluator to rank sentences from four of the systems given the input sentence. Ties are allowed. Table 4 shows the average rat-

	Decoder choice	NO	SO	TO
MT04	Phrase internal	20.7	0.6	21.2
	Phrase external	30.1	43.0	33.1
	Reject	49.2	56.5	45.7
MT05	Phrase internal	21.3	0.7	21.6
	Phrase external	29.5	42.9	31.8
	Reject	49.2	56.4	46.5

Table 5: The reordering choices made based on the three pre-translation reordering approaches for the 20852 and 17195 reorderings proposed by the rules for the MT04 and MT05 test sets. Measured in %.

ings of the systems. This shows the TO scoring to be significantly superior to the other methods ($p < 0.01$ using Wilcoxon signed-rank testing).

6.2.3 MAN vs GDF

Another interesting observation is that reordering rules learned from automatic alignments lead to significantly better translation than rules learned from manual alignment. Due to the much higher quality of the manual alignment, the opposite might be expected. However, this may be just a variant on the observation that alignment improvements (measured against human references) seldom lead to MT improvements (Lopez and Resnik, 2006). The MAN alignments may in fact be better than GDF, but they are most certainly more different in nature from real alignment than the GDF alignments are. As such, the MAN alignments are not as powerful as we would have liked them to be. In our data sets, the GDF rules, seem less specific, and they therefore apply more frequently than the MAN rules. On average, this results in more than 7 times as many possible reordering paths per sentence. This means that the GDF rules supply the decoder with a larger search space, which in turn means more proposed translation hypotheses. This may play a big part in the effect of the rule sets.

6.2.4 Reordering Choices

Table 5 shows the reordering choices made by the approaches in decoding. Most noticeable is that the SO approach is strongly biased against phrase internal reorderings; TO uses more than 30 times as many phrase internal reorderings as SO. In addition, TO is less likely to reject a rule proposed reordering.

The 50 sentences from the manual evaluation are also manually analyzed with regards to reordering. For each reordering in these sentences, the four systems are ranked according to how well the area affected by the reordering is translated. This indicates that the SO approach's bias against phrase internal reorderings may hurt performance. 25% of the time, when SO chooses an external reordering, while the TO approach chooses an internal reordering, the TO approach gets a better translation. Only in 7% of the cases is it the other way around.

Another discovery from the analysis is when TO chooses an internal reordering and NO rejects the reordering. Here, TO leads to a better translation 45% of the time, while NO never outperforms TO. In these cases, either approach has used a phrase to cover the area, but via rule-based motivation, TO has forced a less likely phrase with the correct word order through. This clearly shows that local reordering is not handled sufficiently by phrase internal reordering alone. These need to be controlled too.

6.3 Rule analysis

The rule learning resulted in 61 rules based on manual alignments and 39 based on automatic alignments. Of these, the majority handled the placement of adjectives, while only a few handled the placement of the verb.

A few of the rules that were learned from the manual alignment are shown in table 6. The first two rules handle the placement of the finite verb in Arabic. Rule 16 states that if a finite verb appears in front of a subordinate clause, then it should be moved to sentence initial position with a probability of 68%. Due to the restrictions of sequence lengths, it can only swap across maximally 4 words or a sequence of words that is describable by maximally 3 syntactic phrases. The SBAR condition may help restrict the reordering to finite verbs of the main clause. This rule and its probability goes well with the description given in sections 3, since VSO order is not obligatory. The subject may be unexpressed, or it may appear in front of the verb. This is even more obvious in rule 27, which has a probability of only 43%.

Rules 11 and 1 deal with the inverse ordering of adjectives and nouns. The first is general but uncertain, the second is lexicalized and certain. The reason for the low probability of rule 11 is primarily that many proper names have been mis-tagged by the parser as either JJ or NN, and to a lesser

No	LC	LS	RS	RC	Prob.
16	WORD: <s>		POS: FVF	PS: SBAR	68%
27	WORD: <s>	PS: NP	POS: FVF		43%
11	POS: IN	POS: JJ	POS: NN		46%
1	! POS: JJ	POS: JJ	WORD: president		90%
37	! POS: NN ! POS: JJ	POS: NN	POS: NNS	POS: IN	71%

Table 6: Example rules. ! specifies negative conditions.

extent that the rule should often not apply if the right context is also an NN. Adding the latter restriction narrows the scope of the rule but would have increased the probability to 54%.

Rule 1, on the other hand, has a high probability of 90%. It is only restricted by the condition that the left context should not be an adjective. In these cases, the adjectives should often be moved together, as is the case with *'the south african president'* → الرئيس الجنوب افريقي *Alrŷys Aljnwb Afryqy* where *'south african'* is moved to the right of *'president'*.

Finally, rule 37 handles compound nouns. Here a singular noun is moved to the right of a plural noun, if the right context is a preposition, and the left context is neither an adjective nor a singular noun. This rule handles compound nouns, where the modifying function of the first noun often is hard to distinguish from that of an adjective. The left context restrictions server the same purpose as the left context in rule 1; these should often be moved together with the singular noun. The function of the right context is harder to explain, but without this restriction, the rule would have been much less successful; dropping from a probability of 71% to 51%.

An overall comparison of the rules produced based on the manual and automatic alignments shows no major difference in quality. This is especially interesting in light of the better translation using the GDF rules. It is also very interesting that it seems possible to get as good rules from the GDF as from the MAN alignments. This is a new result compared to Elming (2008), where results on manual alignments only are reported.

7 Conclusion and Future Plans

We have explored the syntactic reordering approach previously presented in (Elming, 2008) within a more distant language pair, English-Arabic. A translation direction that is highly under-represented in MT research, compared to the opposite direction. We achieve significant improvement in translation quality over related approaches, measured by manual as well as automatic evaluations on this task. Thus proving the viability of the approach on distant languages.

We also examined the effect of the alignment method on learning reordering rules. Interestingly, our experiments produced better translation using rules learned from automatic alignments than using rules learned from manual alignments. This is an aspect we want to explore further in the future.

In future work, we would also like to address the morphological complexity of Arabic together with syntax. We plan to consider different segmentations for Arabic and study their interaction with translation and syntactic reordering.

An important aspect of the TO approach is that it uses phrase internal alignments during translation. In the future, we wish to examine the effect their quality has on translation. We are also interested in examining the approach within a standard phrase-based decoder such as Moses (Koehn et al., 2003b) or a hierarchical phrase system (Chiang, 2005).

The idea of training on reordered source language is often connected with pre-translation reordering. The present approach does not employ this strategy, since this is no trivial matter in a non-deterministic, weighted approach. Zhang et al. (2007) proposed an approach that builds on unfolding alignments. This is not an optimal solution, since this may not reflect their rules. Training on both original and reordered data may strengthen the approach, but it would not remedy the problems of the SO approach, since it would still be ignorant of the internal reorderings of a phrase. Nevertheless, it may strengthen the TO approach even further. We also wish to examine this in future work.

References

I. Badr, R. Zbib, and J. Glass. 2008. Segmentation for English-to-Arabic statistical machine translation. In *Proceedings of ACL'08: HLT, Short Papers*, Columbus, OH, USA.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proceedings of ACL'07 Workshop on Statistical Machine Translation*, Prague, Czech Republic.

E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL'00*, Seattle, WA, USA.

D. Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL'05*, Ann Arbor, MI, USA.

W. Cohen. 1996. Learning trees and rules with set-valued features. In *Proceedings of AAAI*, Portland, OR, USA.

M. Collins, P. Koehn, and I. Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of ACL'05*, Ann Arbor, MI, USA.

J. M. Crego and J. B. Mariño. 2007. Syntax-enhanced n-gram-based smt. In *Proceedings of the MT Summit*, Copenhagen, Denmark.

J. Elming. 2008. Syntactic reordering integrated with phrase-based smt. In *Proceedings of the ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, Columbus, OH, USA.

M. Galley, M. Hopkins, K. Knight, and D. Marcu. 2004. What's in a translation rule? In *Proceedings of HLT/NAACL'04*, Boston, MA, USA.

N. Habash and F. Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of HLT-NAACL'06*, New York, NY, USA.

N. Habash, B. Dorr, and C. Monz. 2006. Challenges in Building an Arabic-English GHMT System with SMT Components. In *Proceedings of AMTA'06*, Cambridge, MA, USA.

N. Habash, A. Soudi, and T. Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

N. Habash. 2007. Syntactic preprocessing for statistical machine translation. In *Proceedings of the MT Summit*, Copenhagen, Denmark.

A. Ittycheriah and S. Roukos. 2005. A maximum entropy word aligner for arabic-english machine translation. In *Proceedings of EMNLP*, Vancouver, Canada.

P. Koehn and C. Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings of the Workshop on Statistical Machine Translation at NAACL'06*, New York, NY, USA.

P. Koehn, F. J. Och, and D. Marcu. 2003a. Statistical phrase-based translation. In *Proceedings of NAACL'03*, Edmonton, Canada.

P. Koehn, F. J. Och, and D. Marcu. 2003b. Statistical Phrase-based Translation. In *Proceedings of NAACL'03*, Edmonton, Canada.

P. Koehn, A. Axelrod, A. Birch Mayne, C. Callison-Burch, M. Osborne, and D. Talbot. 2005. Edinburgh system description for the 2005 IWSLT speech translation evaluation. In *International Workshop on Spoken Language Translation 2005 (IWSLT'05)*, Pittsburgh, PA, USA.

P. Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of AMTA'04*, Washington, DC, USA.

Y. Lee. 2004. Morphological Analysis for Statistical Machine Translation. In *Proceedings of HLT-NAACL'04*, Boston, MA, USA.

C. Li, M. Li, D. Zhang, M. Li, M. Zhou, and Y. Guan. 2007. A probabilistic approach to syntax-based reordering for statistical machine translation. In *Proceedings of ACL'07*, Prague, Czech Republic.

A. Lopez and P. Resnik. 2006. Word-based alignment, phrase-based translation: what's the link? In *Proceedings of AMTA'06*, Cambridge, MA, USA.

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

R. Sarikaya and Y. Deng. 2007. Joint morphological-lexical language modeling for machine translation. In *Proceedings of HLT-NAACL'07, Short Papers*, Rochester, NY, USA.

A. Stolcke. 2002. Srilm – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, USA.

C. Wang, M. Collins, and P. Koehn. 2007. Chinese syntactic reordering for statistical machine translation. In *Proceedings of EMNLP-CoNLL*, Prague, Czech Republic.

F. Xia and M. McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING'04*, Geneva, Switzerland.

K. Yamada and K. Knight. 2001. A Syntax-Based Statistical Translation Model. In *Proceedings of ACL'01*, Toulouse, France.

R. Zens, F. J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In M. Jarke, J. Koehler, and G. Lakemeyer, editors, *KI - 2002: Advances in Artificial Intelligence. 25. Annual German Conference on AI*. Springer Verlag.

Y. Zhang, S. Vogel, and A. Waibel. 2004. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Y. Zhang, R. Zens, and H. Ney. 2007. Improved chunk-level reordering for statistical machine translation. In *Proceedings of the IWSLT*, Trento, Italy.