Mach Translat (2012) 26:105–120
DOI 10.1007/s10590-011-9112-y

Improved Arabic-to-English statistical machine translation by reordering post-verbal subjects for word alignment

Marine Carpuat · Yuval Marton · Nizar Habash

Received: 5 July 2010 / Accepted: 27 August 2011 / Published online: 30 October 2011

Abstract We study challenges raised by the order of Arabic verbs and their subjects in statistical machine translation (SMT). We show that the boundaries of post-verbal subjects (VS) are hard to detect accurately, even with a state-of-the-art Arabic dependency parser. In addition, VS constructions have highly ambiguous reordering patterns when translated to English, and these patterns are very different for matrix (main clause) VS and non-matrix (subordinate clause) VS. Based on this analysis, we propose a novel method for leveraging VS information in SMT: we reorder VS constructions into pre-verbal (SV) order for word alignment. Unlike previous approaches to source-side reordering, phrase extraction and decoding are performed using the original Arabic word order. This strategy significantly improves BLEU and TER scores, even on a strong large-scale baseline. Limiting reordering to matrix VS yields further improvements.

Keywords Statistical machine translation · Reordering · VS · Post-verbal subjects · Matrix subject · Subject detection · Word alignment · Dependency parsing

M. Carpuat (✉)
National Research Council, 283 Alexandre-Taché Boulevard, Building CRTL,
Gatineau, QC J8X 3X7, Canada
e-mail: Marine.Carpuat@cnrc-nrc.gc.ca; marine@ccls.columbia.edu

Y. Marton
IBM T. J. Watson Research Center, Kitchawan Road/Route 134,
Yorktown Heights, NY 10598, USA
e-mail: yymarton@us.ibm.com

N. Habash
Columbia University Center for Computational Learning Systems, 475 Riverside Drive MC 7717,
New York, NY 10115, USA
e-mail: habash@ccls.columbia.edu

1 Introduction

Arabic verbs and their subjects pose many challenges to Arabic–English statistical machine translation (SMT). Arabic subjects can occur in pre-verbal (SV), post-verbal (VS) or pro-dropped constructions (VNS—verbs with no explicit subject), while English is primarily SV; Arabic gender and number agreement rules differ in SV and VS orders; subjects can be long, in particular when they include recursive possessive constructions. These variations make it particularly hard to automatically detect, word align and translate Arabic verbs and subjects correctly.

In this article, we first attempt to get a better understanding of translation patterns for Arabic verbs and their subjects, particularly VS constructions, by studying their occurrence and reordering patterns in a hand-aligned Arabic–English parallel treebank. Our analysis shows that VS reordering rules are not straightforward and that SMT should therefore benefit from direct modeling of Arabic verb and subject translation.

We then turn to detecting these constructions automatically, which is a challenging task in itself. Using a state-of-the-art Arabic dependency parser, we show that VS constructions and their exact boundaries are hard to identify accurately. Given this noise in VS detection, existing strategies for source-side reordering (e.g., Xia and McCord 2004; Collins et al. 2005; Wang et al. 2007) or using dependency parses as cohesion constraints in decoding (Cherry 2008; Bach et al. 2009) are not effective at this stage. While these approaches have been successful for language pairs such as German–English for which syntactic parsers are more developed and relevant reordering patterns might be less ambiguous, their impact potential on Arabic–English translation is still unclear.

We therefore focus on VS constructions, and propose a strategy to benefit from their noisy detection in SMT for the *word alignment stage only*. We reorder phrases detected as VS constructions into an SV order. Unlike in previous syntactic reordering approaches, subjects are moved back to the original VS word order before phrase-extraction. For phrase extraction, weight optimization and decoding, we use the original (non-reordered) text. While this strategy does not address the important problem of reordering at decoding time, it successfully leverages subject span information, and yields significant improvements on BLEU and TER on top of strong medium and large-scale phrase-based SMT baselines. We further show that limiting reordering to matrix VS subjects yields additional gains on both medium- and large-scale settings. This simple but crucial modification of the reordering rule is motivated by two observations:

- First, we show that matrix and non matrix VS have very different reordering patterns. Using a manually word-aligned Arabic–English corpus, we discover that while most matrix VS constructions are translated in inverted order (SV), non-matrix VS constructions are inverted in only half the cases.
- Second, while detecting verbs and their subjects is a hard task, our syntactic parser detects VS constructions better in matrix than in non-matrix clauses. Reordering only matrix VS therefore introduces less noise due to incorrect parses.

This article draws together our work in Carpuat et al. (2010a) and Carpuat et al. (2010b), and extends it with further analysis of the impact of subject reordering on word alignment quality. To the best of our knowledge, the only other attempt at explicitly modeling Arabic subjects for translation failed to improve phrase-based SMT (Green et al. 2009). In contrast, Bisazza and Federico (2010) who proposed a similarly motivated reordering method for the entire SMT pipeline, rely on Arabic base-phrase chunks and sidestep the issue of subject detection.

In Sect. 2, we present a discussion of the various challenges of processing the Arabic verb-and-subject constructions. Section 3 outlines our approach and presents an evaluation of it. Section 4 presents a deeper analysis of the produced alignments. Finally Sects. 5 and 6 present a discussion of related work and conclude this article, respectively.

2 Processing challenges of the Arabic verb-and-subject constructions

In this section, we discuss relevant linguistic facts on the Arabic verb-and-subject constructions. Then we present two sets of analyses for reordering in MT and automatic parsing.

2.1 Relevant linguistic facts

Arabic is a morpho-syntactically complex language with many differences from English. We describe here two linguistic features of Arabic that are relevant to Arabic–English translation and how we handle them: Arabic's complex morphology, and VS order.[1]

First, Arabic words are morphologically complex containing clitics whose translations are represented separately in English and sometimes in a different order. For instance, possessive pronominal enclitics are attached to the noun they modify in Arabic but their translation precedes the English translation of the noun: كتاب+ه *kitAbu+hu*[2] 'book+his → his book'. Other clitics include the definite article +ال*Al*+ 'the', the conjunction +و*w*+ 'and' and the preposition +ل*l*+ 'of/for', among others. Separating some of these clitics has been shown to help SMT (Habash and Sadat 2006). In this article we do not investigate which clitics to separate, but instead we use the Penn Arabic Treebank (PATB) (Maamouri et al. 2004) tokenization scheme which splits all clitics except for the definite article +ال *Al*+ (see example in Fig. 1). We tokenize our data using the Morphological Analysis and Disambiguation for Arabic (MADA+TOKAN v. 3.1) toolkit (Habash and Rambow 2005; Habash 2007a,b; Roth et al. 2008), for both parser training purposes and SMT (word alignment, phrase extraction, and decoding).

[1] Other cases of Arabic constructions undergo complex reordering too when translated to English, e.g., Noun–Noun (Idafa) and Noun–Adjective constructs. They are usually easily handled in phrase-based SMT system using a relatively short phrase size and local distortion. As such, we do not offer any solutions other than the basic phrase-based MT setup.

[2] All Arabic transliterations are presented in the HSB transliteration scheme (Habash et al. 2007).

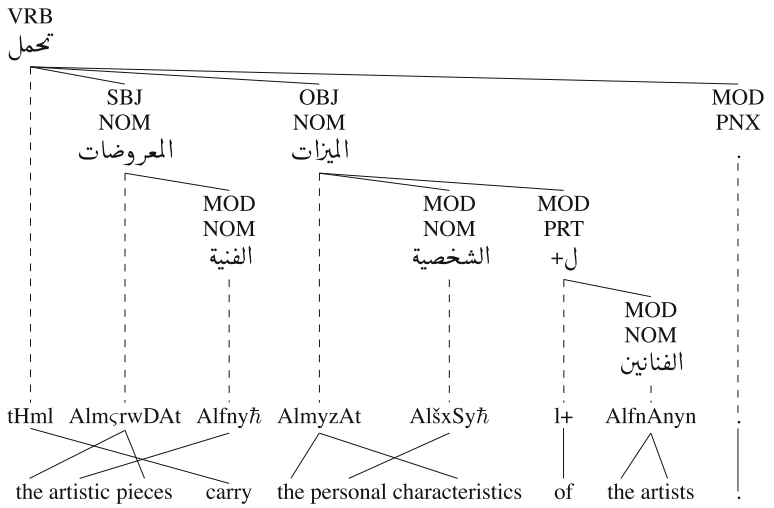

Fig. 1 A pair of word-aligned Arabic and English sentences. The Arabic syntactic dependency representation is in CATiB style annotation

[V اعلن**]** [SBJ المنسق العام ل+ مشروع السكة الحديد بين دول مجلس التعاون الخليجي] *[OBJ* ان *...]*

[V Açln] [SBJ Almnsq AlçAm l+ mšrwç Alskħ AlHdyd byn dwl mjls AltçAwn Alxlyjy] *[OBJ An ...]*

[SBJ The general coordinator of the railroad project among the countries of the Gulf Cooperation Council] **[V announced]** *[OBJ that ...]*

Fig. 2 An example of long distance reordering of Arabic VS order into English SV order

Second, the subject in Arabic VS constructions may be: (a) pro-dropped (conjugated verb), (b) pre-verbal (SV), or (c.) post-verbal (VS). Each situation comes with its own morphosyntactic restrictions. Generally, verbs agree with subjects in person, gender and number in SV order, but only in person and gender in VS order. From the point of reordering, the case of VS order is the most interesting in the context of translation to English (see Fig. 1). For small noun phrases (NP), phrase-based SMT 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. Figure 2 illustrates this point: boldface and italics are used to mark the verb and subordinating conjunction that surround the subject NP (11 tokens) in Arabic and what they map to in English, respectively. Additionally, since Arabic is a pro-drop language, we cannot "blindly" move the NP following the verb, since it can be the object of that verb. A mistaken identification of the subject boundaries can lead to moving part of the subject before the verb and keeping the rest after, which is likely to hurt word alignment. These observations illustrate the importance of having a suitable syntactic analyzer that can not only identify the boundaries of NP (and other potential subjects) but also assign them the correct relation to the correct verb in the sentence. For more information on Arabic morphology and syntax, see Habash (2010).

Table 1 How are Arabic SV and VS translated in the manually word-aligned Arabic–English parallel treebank?

	Gold reordering	All verbs	%	Matrix	%	Non-matrix	%
SV	Monotone	2588	**98.2**	625	**98.4**	1963	**98**
SV	Inverted	15	0.5	0	0	15	0.7
SV	Overlap	35	1.3	10	1.6	25	1.3
SV	Total	2638	100	635	100	2103	100
VS	Monotone	1700	27.3	421	13.6	1279	40.8
VS	Inverted	4033	**64.7**	2524	**81.4**	1509	**48.1**
VS	Overlap	502	8	154	5	348	11.1
VS	Total	6235	100	3099	100	3136	100

We check whether V and S are translated in a "monotone" or "inverted" order for all VS and SV constructions. "Overlap" represents instances where translations of the Arabic verb and subject have some English words in common, and are not monotone nor inverted
Bold characters indicate the most frequent reordering operation for each type of construction

2.2 Reordering rules for gold Arabic VS are not deterministic

We use the manually word-aligned parallel Arabic–English Treebank (LDC2009E82) to study how Arabic VS constructions are translated into English by humans. Given the gold Arabic syntactic parses and the manual Arabic–English word alignments, we can determine the gold reorderings for SV and VS constructions. We extract VS representations from the gold constituent parses by deterministic conversion to a simplified dependency structure, CATiB (Habash and Roth 2009) (see Sect. 2).

We then check whether the English translations of the Arabic verb and the Arabic subject occur in the same order as in Arabic (monotone) or not (inverted). Table 1 summarizes the reordering patterns for each category. As expected, 98% of Arabic SV are translated in a monotone order in English. For VS constructions, the picture is surprisingly more complex. The majority of Arabic VS are reordered into English (as in the examples in Figs. 1 and 2), but 27% are translated in a monotone order. A common case of monotone translation order in English is the case of VS constructions in subordinating clauses mapping into English passive constructions, e.g., the Arabic الكتاب الذي كتب +ه علي *AlktAb Alðy ktb* + *hς ly* 'the book that Ali wrote' may be translated as 'the book written by Ali'.

We manually inspected the data and found that many monotone VS occur in subordinate clauses. This suggested that distinguishing between matrix (main) versus non-matrix (subordinate) subjects might provide additional insights. Interestingly, VS in matrix clauses are reordered more often (81%) than non-matrix VS (48%). The monotone VS translations are mostly explained by changes to passive voice or to non-verbal constructions (such as nominalization) in the English translation.

In addition, Table 1 shows that subjects occur more frequently in VS order (70%) than in SV order (30%). These numbers do not include pro-dropped ("null subject") constructions, since generating correct translations for them is not a reordering issue.

2.3 Arabic VS constructions are hard to identify

Before turning to translation, we need to identify Arabic subjects, their spans, and the verbs they attach to (and potentially reorder with). We employ a dependency parser for this task and show that it detects verbs and their subjects with an F-score of 74%.

Most statistical syntactic parsers used in SMT are constituency parsers (Bikel 2004; Manning and Schuetze 1999), and do not typically mark subject relations explicitly. In contrast, in Carpuat et al. (2010a) and Carpuat et al. (2010b) we employ a dependency parser—MaltParser with the Nivre "eager" algorithm (Nivre 2003; Nivre et al. 2006) [3]—as follows: We train the parser on the training portion of the University of Pennsylvania Arabic Treebank (PATB) part 3 (v3.1) (Maamouri et al. 2004), with the dev/test split defined by Zitouni et al. (2006). As in the Columbia Arabic Treebank (CATiB) (Habash and Roth 2009), we convert the PATB annotation to a simplified format with 8 dependency relations and 6 POS tags, to gain higher POS prediction accuracy. We then extend it to a set of 44 tags using regular expressions of the basic POS and a linguistically motivated set of affixes of the normalized surface word forms. (We normalize Alif Maqsura to Ya, and Hamzated Alifs to bare Alif, as is commonly done in Arabic SMT). This parsing model is described in more details in Marton et al. (2010): it is essentially the CATIBEX (extended CATiB POS tag) baseline model. Further discussion and subsequent work on the parsing models can be found there. Evaluated on the development section of PATB 3v3.1, our parsing model achieves an overall labeled attachment score of 79.25%, using MADA predicted (non-gold) POS tags.

In this work, we are specifically interested in (a) detection of subjects (with their correct span) in constructions with verbs, and (b) detection of the verb that governs each subject and which determines where the subject should move *to* in the translation to the target language (English). Hence, we argue that combined detection statistics of constructions of both verbs and their subjects (VATS hereafter) are more telling, when evaluating parsing quality for reordering.[4] Table 2 includes overall precision/recall/F-score statistics for all VATS and for each type of verbal construction (VS, SV and VNS) regardless of matricity and also for matrix/non-matrix conditions. VNS refers to verbs with no explicit separate subject token (a.k.a. pro-drop or null-subject verbs).

[3] Nivre (2008) reports that non-projective and pseudo-projective algorithms outperform the "eager" projective algorithm in MaltParser; however, our converted training data contain no non-projective dependencies, so there was no point in using these algorithms. The Nivre "standard" algorithm is also reported to do better on Arabic, but in a preliminary experimentation, it did slightly worse than the "eager" one. This could be due to high percentage of right branching (left headed structures) in Arabic, an observation already noted in Nivre (2008).

[4] We divert from the CATiB representation in that a non-matrix subject of a pseudo verb (إن وأخواتها) is treated as a subject of the verb that is under the same pseudo verb. This treatment of said subjects is comparable to the PATBs. Note also that a matrix subject or verb that is mis-identified as non-matrix, or vice versa, does not get credit in our scoring; neither does a partially detected span.

Table 2 Subject and verb detection precision, recall and F-scores

	All (matricity-insensitive)				Matrix				Non-matrix			
	%	P	R	F	%	P	R	F	%	P	R	F
VATS	100	73.84	74.37	74.11	32	65.06	68.01	66.50	68	75.91	75.06	75.48
VS	37	66.62	59.41	62.81	57	68.1	62.59	65.25	28	62.18	53.81	57.69
SV	18	86.75	61.07	71.68	18	81.82	53.33	64.57	19	85.98	62.59	72.44
VNS	44	76.32	92.04	83.45	25	56.37	90.31	69.41	53	79.21	90.02	84.27

VATS: (All) Verbs and their subjects, regardless of subject form or construction, *VS*, *SV* verb-subject and subject-verb constructions, respectively, *VNS* verbs with null subjects (having no separate token for subject). In the VATS row, the % column cells are for percentage of all VATS; however, the other % column cells are for percentage of all VATS in the same matricity condition

VS refers only to verbs (whether PV or IV) with subjects that are NP. NP with NULL heads are deleted in CATiB and verbs with such subjects are marked as VNS.

- Overall, identifying VATS is hard, with 74% F-score. Matrix VATS are much harder to detect—almost 9% absolute lower than non-matrix VATS. This difference is partly explained by errors in identifying whether a verb is a matrix verb or not. In addition, recall that these scores reflect both errors in detecting subject spans and in identifying the verb the subject is attached to.
- VS constructions, our main focus for reordering, have the lowest F-score of all constructions: 63%. However, VS constructions are detected with a higher F-score in the harder matrix condition than in the non-matrix condition. This differs from the other two types of constructions which fare better in the non-matrix condition.
- The low precision of the matrix VNS condition (56.37% in Table 2) reveals another source of errors. A VNS construction (null-subject verb) is correct if the verb is tagged correctly and has a null-subject. Since matrix verbs are tagged with high precision (almost 93%)[5], most errors in matrix VNS detection are due to unidentified subjects, i.e. VS and SV constructions that are incorrectly identified as VNS.

To the best of our knowledge, we cannot directly compare these numbers to any previously published work. For instance, the Stanford Arabic parser (Green and Manning 2010) is a constituency parser that does not identify VS. The closest basis for comparison is work by Green et al. (2009) (see Sect. 5), who propose a VS detection technique that bypasses syntactic parsing. Instead, they use conditional random fields to detect only maximal (non-nested) subjects of verb-initial clauses. They report 65.9% precision and 61.3% F-score, but use a different training/test split of the PATB data (parts 1, 2 and 3).

[5] Note that this evaluation starts with gold tokenization.

3 Reordering Arabic VS for SMT word alignment

3.1 Approach

Based on these analyses, we propose a new method to help phrase-based SMT systems deal with Arabic–English word order differences due to VS constructions. As in related work on syntactic reordering by preprocessing, our method aims to make Arabic and English word order closer to each other by reordering Arabic VS constructions into SV. However, unlike in previous work, the reordered Arabic sentences are used only for word alignment. Phrase translation extraction and decoding are performed on the original Arabic word order. Given a parallel sentence (a, e), we proceed as follows:

(1) automatically tag VS constructions in a
(2) generate new sentence $a' = reorder(a)$ by reordering Arabic VS into SV
(3) get word alignment wa' on new sentence pair (a', e)
(4) using mapping from a to a', get corresponding word alignment $wa = unreorder(wa')$ for the original sentence pair (a, e)

Based on the analysis of gold reordering patterns and our automatic subject detection tools, we also introduce a simple but crucial variation to step 2, where reordering is limited to *matrix* VS constructions.

Reordering Arabic VS attempts to make the bitext more monotone and therefore easier to explain by the alignment model. Commonly used alignment models have weak reordering models and prefer monotone alignments. For instance, in IBM models 2 and 3 (Brown et al. 1993), the distortion model is only based on absolute word positions in the source and target sentence, while IBM-4 and HMM models (Vogel et al. 1996) use relative word positions and condition distortion on the alignment of the previous word.

Limiting reordering to alignment prevents the system from learning translation rules on incorrect word orders introduced either by incorrect VS detection, or by incorrect reordering of a possibly correctly detected VS. Experiments on an earlier version of the large-scale SMT system described in Sect. 3 showed that such errors are common since forcing reordering of VS constructions at training and test time does not have a consistent impact on translation quality.[6] These results are consistent with recent reports that forcing reordering throughout the translation process has a mixed impact on translation quality even for German–English SMT (Howlett and Dras 2011). Taken together, these results suggest that integrating VS reordering in decoding requires more sophisticated models, and we plan to address this in future work (Andreas et al. 2011, for follow-up work on these ideas). In this article, we choose to limit our experiments to reordering for alignment, and we do not directly address the weaknesses of the reordering model used at decoding time. For instance, when used with a phrase-based SMT decoder, our approach can improve the phrase-table and the lexicalized reordering model, but it has no direct impact on the ability of the decoder to handle long-range reorderings.

[6] For instance, on the NIST MT08-NW test set, reordering all VS constructions improved TER slightly from 44.34 to 44.03, while BLEU score decreased from 49.21 to 49.09. Reordering matrix VS only degraded TER to 45.76, and BLEU to 46.86.

3.2 SMT evaluation set-up

We use the open-source Moses toolkit (Koehn et al. 2007) to build two phrase-based SMT systems trained on two different data conditions:

(1) *Medium-scale* the bitext consists of 12M words on the Arabic side (LDC2007E103). The language model is trained on the English side of the large bitext.
(2) *Large-scale* the bitext consists of several newswire LDC corpora, and has 64M words on the Arabic side. The language model is trained on the English side of the bitext augmented with Gigaword data.

For both systems, the parallel corpus is word-aligned using GIZA++ (Och and Ney 2003), which sequentially learns word alignments for the IBM1, HMM, IBM3 and IBM4 models. The resulting alignments in both translation directions are intersected and augmented using the grow-diag-final-and heuristic (Koehn et al. 2007). Phrase translations of up to 10 words are extracted in the Moses phrase-table, and filtered using statistical significance testing (Johnson et al. 2007). We use a 5-gram language model with modified Kneser–Ney smoothing, and lexicalized reordering (monotone, swap and discontinuous orientations, in both translation directions). The weights for the five phrase-table features, six lexicalized distortion features and the language model scores are tuned to maximize BLEU on the NIST MT06 test set. The English data is tokenized using simple punctuation-based rules. The Arabic side is segmented according to the Arabic Treebank v3.1 tokenization scheme using the MADA+TOKAN morphological analyzer and tokenizer (Habash and Rambow 2005; Habash 2007a,b; Roth et al. 2008). MADA-produced Arabic lemmas are used for word alignment. The dependency parser described in Sect. 2 is applied to the entire Arabic training data.

3.3 Results: VS reordering significantly improves BLEU and TER

We evaluate the performance of all reordering for alignment variants on five of the NIST Arabic–English test sets. As can be seen in Table 3, on a large test set of more than 4,440 sentences, reordering matrix VS remarkably yields statistically significant improvements in BLEU (Papineni et al. 2002) and TER (Snover et al. 2006) over both baseline SMT systems at the 99% confidence level (Koehn 2004). In addition, restricting reordering to matrix VS also yields better scores than reordering all VS constructions. Results per test set are reported in Table 4. It is worth noting that consistent improvements are obtained even on the large-scale system, and that both medium and large-scale baselines are strong full-fledged systems with distortion and lexicalized reordering models, as well as large 5-gram language models.

4 Analysis of reordered alignments

We next present three sets of analyses comparing the reordered alignments to manual alignments, baseline alignments, and large-data oracle alignments.

Table 3 Evaluation on all test sets: on the total of 4,432 test sentences, improvements are highly statistically significant (99% level using bootstrap resampling (Koehn 2004))

System	BLEU r4n4 (%)	TER (%)
Medium baseline	44.35	48.34
+ All VS reordering	44.65 (+0.3)	47.78 (−0.56)
+ Matrix VS reordering	44.96 (+0.61)	47.52 (−0.82)
Large baseline	51.45	42.45
+ All VS reordering	51.70 (+0.25)	42.21 (−0.24)
+ Matrix VS reordering	51.80 (+0.35)	42.11 (−0.34)

Table 4 VS reordering improves BLEU and TER scores in almost all test conditions on 5 test sets, 2 metrics, and 2 MT systems

Test set	MT03	MT04	MT05	MT08nw	MT08wb
BLEU r4n4 (%)					
Medium baseline	45.95	44.94	48.05	44.86	32.05
+ All VS reordering	46.33	45.03	48.69	45.06	31.96
+ Matrix VS reordering	46.79	45.28	49.11	45.19	31.98
Large baseline	52.30	52.45	54.66	52.6	39.22
+ All VS reordering	52.63	52.34	55.29	52.85	39.87
+ Matrix VS reordering	52.88	52.42	55.29	52.98	40.01
TER (%)					
Medium baseline	48.764	46.452	44.998	47.744	58.022
+ All VS reordering	47.878	46.153	44.140	47.284	57.339
+ Matrix VS reordering	48.311	46.103	44.286	47.115	57.304
Large baseline	43.327	40.414	39.154	41.807	52.049
+ All VS reordering	42.778	40.338	38.747	41.364	52.005
+ Matrix VS reordering	42.951	40.398	38.747	41.513	51.859

4.1 Comparison of reordered alignments against manual word alignments

How do the gains in BLEU and TER relate to the changes in word alignment introduced by matrix VS reordering?

Since manual word alignments are available in the English–Arabic parallel treebank, we can evaluate our word alignment strategies intrinsically by computing their Alignment Error Rate with respect to the manual alignments. Our test set comprised 4,630 sentences from the broadcast news section of the parallel treebank. Note that we did not include the newswire section in the evaluation since it was used as training data for our dependency parser.

Using the medium-scale models, the alignments obtained with the baseline system and matrix VS reordering yield very close error rates (58.42 and 58.45 respectively), while reordering all VS yields a slightly higher error rate (58.58). Since most of the words are unaffected by our reordering strategy, it is not surprising that the difference in

Table 5 Comparison of alignment links learned with and without reordering: in columns 3–6, the number in row *i* and column *j* represents the percentage of alignment links in system *i* that are identical to alignment links in system *j* on a sample of 15k sentence pairs

System	No. links	Med baseline (%)	+ All VS (%)	+ Matrix VS (%)	Large-data oracle (%)
Medium baseline	330255	100.00	87.64	43.28	66.05
+ All VS	330255	87.64	100.00	67.75	58.49
+ Matrix VS	326625	75.51	67.00	100.00	66.35

error rates are small. These numbers essentially suggest that the automatic alignments are more similar to each other than to the manual alignment. However, we note that the reordering strategies that improve BLEU and TER do not help the alignment error rate. These results are consistent with previous work, showing that intrinsic evaluation of word alignment quality against manually created references does not correlate well with translation quality (see Lopez and Resnik 2006, for an overview).

4.2 Comparison of reordered alignments against baseline alignments

We compare the word alignment links learned by the different versions of the system. For this comparison, we use a subset of the medium-scale bitext of about 15,000 sentence pairs, and compute the number of common alignment links that are identical for each pair of alignment methods.

Table 5 shows that reordering matrix VS yields slightly fewer alignment links than both the baseline and the system that reordered all VS. Columns 3–5 shows that the word alignments learned with reordered matrix VS are quite different from all others: only 43% of these links are also learned by the baseline system, while more than 75% of the baseline links are covered with VS matrix reordering.

4.3 Comparison of reordered alignments against large-data oracle alignments

Finally, we compare the reordered alignments with large-data oracle alignments, that, unlike manual alignments, provably improve translation quality.

Since our large-scale baseline is trained on a superset of the medium-scale bitext, we use the word alignments learned for the large-scale baseline as large-data oracle alignments for the medium-scale system. As described in Sect. 3, the large-scale alignments are learned on a bitext that is more than five times larger than the medium-scale bitext.

Here, the large-scale alignments are viewed as a reference for comparison with alignment links obtained with all medium-scale systems. We use the same sample of sentences as in Sect. 4 and report the results in Table 4, Column 6: of all three medium-scale conditions, the matrix VS reordering strategy yields the highest percentage of common links with the large-scale alignments.

In addition, since, unlike manual alignments, the large-data oracle alignments are available for the entire training bitext, we can directly quantify their impact on end-to-end translation quality. We therefore build a fourth medium-scale system using the

large-data oracle alignments: this system is trained using alignment links learned on the large bitext, but only for the subset of the bitext that matches the medium-scale data condition.[7] This oracle system improves the medium-scale baseline by +1.37 BLEU and −1.6 TER on the concatenated test sets. Comparing these improvements in BLEU and TER with those obtained in Table 3 shows that the gains obtained with VS reordering are quite significant: without using any additional SMT training data, our matrix VS reordering technique interestingly yields 44% of the gain in BLEU and 51% of the gain in TER obtained with the large-data oracle.

5 Related work

To the best of our knowledge, the only other approach to detecting and using Arabic VS constructions for SMT is that of Green et al. (2009), which failed to improve Arabic–English SMT. Instead of directly modeling VS reordering, subject span information was used to encourage a phrase-based SMT decoder to use phrasal translations that do not break subject boundaries. Matrix and non-matrix subjects were not treated differently. In addition, their VS detection model is very different from ours, since it bypasses full syntactic parsing, but similarly produces noisy subject boundaries, especially at the "right edge". They report 65.9% precision and 61.3% F-score only detecting maximal (non-nested) subjects of verb-initial clauses (most comparable to our VS condition) using a different training/test split of the PATB (parts 1, 2 and 3) data. Both approaches use simplified POS tags , and various linguistic relations, such as the N–N construct (*Idafa*). However, while they use a generally flat (non-hierarchical) notation, trained with conditional random fields, we rely on hierarchical representations from dependency parsing, allowing us coverage of non-maximal subjects as well, in addition to matricity identification.

Syntactically motivated reordering for phrase-based SMT has been more successful on other language pairs than Arabic–English, perhaps due to more accurate parsers and less ambiguous reordering patterns than for Arabic VS. For instance, Collins et al. (2005) apply six manually defined transformations to German parse trees which yield an improvement of 1.6 BLEU on the Europarl German–English translation task. Recent results show however that such improvements are reachable only in specific training conditions, in particular when training and test data are very close to each other (Howlett and Dras 2011). Xia and McCord (2004) learn reordering rules for French to English translations, which arguably presents less syntactic distortion than Arabic–English. Zhang et al. (2007) limit reordering to decoding for Chinese–English SMT using a lattice representation. Cherry (2008) uses dependency parses as cohesion constraints in decoding for French–English SMT. In addition, note that even syntax-aware SMT models often do not directly capture subject information, as they typically rely on phrase-structure representations (Marton and Resnik 2008 *inter alia*.)

[7] Of course, this unusual training regimen is an artificial experiment setting, which we only use here for the purpose of analysis. For a real translation task, there is no reason to use more data for word alignment than for the rest of the SMT training pipeline.

For Arabic–English phrase-based SMT, the impact of syntactic reordering as preprocessing is less clear. Habash (2007b) shows that syntactic reordering rules targeting Arabic–English word order differences help BLEU compared to phrase-based SMT limited to monotonic decoding, but improvements do not hold with distortion. Learning reordering rules has given positive results when using POS and shallow syntax in a ngram-based SMT system (Crego and Habash 2008). Recently, Bisazza and Federico (2010) have reported promising results with lattice decoding for reordering clause-initial verbs in phrase-based SMT. Interestingly, they bypass the difficult problem of subject detection by automatically learning reordering rules based on base phrase chunks, rather than explicitly identifying subjects and their reordering patterns. It would be interesting to combine lattice decoding with our word alignment strategy, and compare the impact of full subject detection with that of shallow syntactic analysis.[8]

Most previous syntax-aware word alignment models were specifically designed for syntax-based SMT systems. These models are often bootstrapped from existing word alignments, and could therefore benefit from our VS reordering approach. For instance, Fossum et al. (2008), report improvements ranging from 0.1 to 0.5 BLEU on Arabic translation by learning to delete alignment links if they degrade their syntax-based translation system. Departing from commonly-used alignment strategies, Hermjakob (2009) aligns Arabic and English content words using pointwise mutual information, and in this process indirectly uses English sentences reordered into VS order to collect cooccurrence counts. The approach outperforms GIZA++ on a small scale translation task, but the impact of reordering alone is not evaluated.

6 Conclusion

We presented a novel method for improving overall SMT quality using noisy syntactic dependency parses: we use these parses to reorder VS constructions into SV order for word alignment only. This approach increases word alignment coverage and significantly improves BLEU and TER scores on two strong SMT baselines.

In addition, we showed that matrix VS constructions deserve special attention in Arabic-to-English translation. While most matrix VS constructions are translated in inverted order (SV), non-matrix (subordinate clause) VS constructions are inverted in only half the cases. Moreover, matrix VS construction spans are better detected than non-matrix VS. This suggest that it is not advisable to work under the naïve assumption that all Arabic VS constructions should be translated to English SV. Based on this observation, we refine the reordering rule applied to word alignments with a simple but crucial change: instead of reordering all Arabic VS constructions, we limit reordering to matrix VS. This approach yields further improvements in translation quality.

Acknowledgements The authors would like to thank Mona Diab, Owen Rambow, Ryan Roth, Kristen Parton and Joakim Nivre for helpful discussions and assistance. The first two authors started

[8] During the period of finalizing this article, Andreas et al. (2011) presented a study combining these ideas and extending the work presented here by introducing multiple fuzzy reorderings during decoding.

work for this article while at Columbia University, NY. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under GALE Contract No. HR0011-08-C-0110. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency.

References

Andreas J, Habash N, Rambow O (2011) Fuzzy syntactic reordering for phrase-based statistical machine translation. In: Proceedings of the workshop on statistical machine translation at the conference on empirical methods for natural language processing (EMNLP), Edinburgh, Scotland, UK

Bach N, Vogel S, Cherry C (2009) Cohesive constraints in a beam search phrase-based decoder. In: Proceedings of the 10th meeting of the North American chapter of the Association for Computational Linguistics, Companion volume: Short papers, Boulder, Colorado, pp 1–4

Bikel DM (2004) Intricacies of Collins' parsing model. Comput Linguist 30(4):479–511

Bisazza A, Federico M (2010) Chunk-based verb reordering in VSO sentences for Arabic-English statistical machine translation. In: Proceedings of the joint fifth workshop on statistical machine translation and MetricsMATR, Uppsala, Sweden, pp 235–243

Brown PE, Pietra VJD, Pietra SAD, Mercer RL (1993) The mathematics of statistical machine translation: parameter estimation. Comput Linguist 19(2):263–312

Carpuat M, Marton Y, Habash N (2010a) Explorations in subject-verb reordering for Arabic-English statistical machine translation. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL 2010), short papers, Los Angeles, CA, pp 178–183

Carpuat M, Marton Y, Habash N (2010b) Reordering matrix post-verbal subjects for Arabic-to-English SMT. In: Proceedings of the 17th conference sur le Traitement des Langues Naturelles (TALN), Montreal, Canada

Cherry C (2008) Cohesive phrase-based decoding for statistical machine translation. In: Proceedings of ACL'08, Columbus, Ohio, pp 72–80

Collins M, Koehn P, Kucerova I (2005) Clause restructuring for statistical machine translation. In: Proceedings of ACL 2005 (meeting of the Association for Computational Linguistics), Ann Arbor, Michigan pp 531–540

Crego JM, Habash N (2008) Using shallow syntax information to improve word alignment and reordering for SMT. In: Proceedings of the third workshop on statistical machine translation, Columbus, Ohio, pp 53–61

Fossum V, Knight K, Abney S (2008) Using syntax to improve word alignment precision for syntax-based machine translation. In: StatMT '08: proceedings of the third workshop on statistical machine translation, Waikiki, Hawai'i, pp 44–52

Green S, Manning CD (2010) Better Arabic parsing: baselines, evaluations, and analysis. In: Proceedings of the 23rd international conference on computational linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, pp 394–402

Green S, Sathi C, Manning CD (2009) NP subject detection in verb-initial Arabic clauses. In: Proceedings of the third workshop on computational approaches to Arabic script-based languages (CAASL3), Ottawa, Ontario

Habash N (2007a) Arabic morphological representations for machine translation. In: van den Bosch A, Soudi A (eds). Arabic computational morphology: knowledge-based and empirical methods. Springer, Dordrecht

Habash N (2007b) Syntactic preprocessing for statistical machine translation. In: Proceedings of the 11th machine translation summit (MT-Summit), Copenhagen, Denmark

Habash N (2010) Introduction to Arabic natural language processing. Morgan and Claypool Publishers, San Rafael

Habash N, Rambow O (2005) Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05), Ann Arbor, Michigan, pp 573–580

Habash N, Roth R (2009) CATiB: The Columbia Arabic treebank. In: Proceedings of the ACL-IJCNLP 2009 conference short papers, Suntec, Singapore, pp 221–224

Habash N, Sadat F (2006) Arabic preprocessing schemes for statistical machine translation. In: Proceedings of the 7th meeting of the North American chapter of the Association for Computational Linguistics, Companion volume: short papers, New York City, USA, pp 49–52

Habash N, Soudi A, Buckwalter T (2007) On Arabic transliteration. In: van den Bosch A, Soudi A (eds). Arabic computational morphology: knowledge-based and empirical methods. Springer, Berlin

Hermjakob U (2009) Improved word alignment with statistics and linguistic heuristics. In: Proceedings of the conference on empirical methods in natural language processing, Singapore, pp 229–237

Howlett S, Dras M (2011) Clause restructuring for SMT not absolutely helpful. In: Proceedings of the 49th annual meeting of the Association for Computational Linguistics, Portland, Oregon

Johnson H, Martin J, Foster G, Kuhn R (2007) Improving translation quality by discarding most of the phrasetable. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, Czech Republic, pp 967–975

Koehn P (2004) Statistical significance tests for machine translation evaluation. In: Proceedings of the 2004 conference on empirical methods in natural language processing (EMNLP-2004), Barcelona, Spain, pp 388–395

Koehn P, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C, Zens R, Dyer C, Bojar O, Constantin A, Herbst E (2007) Moses: open source toolkit for statistical machine translation. In: Annual meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, pp 177–180

Lopez A, Resnik P (2006) Word-based alignment, phrase-based translation: What's the link? In: Proceedings of the 7th conference of the Association for Machine Translation in the Americas, Cambridge, Massachusets pp 90–99

Maamouri M, Bies A, Buckwalter T, Mekki W (2004) The Penn Arabic treebank: building a large-scale annotated Arabic corpus. In: NEMLAR conference on Arabic language resources and tools, Cairo, Egypt, pp 102–109

Manning CD, Schuetze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, MA

Marton Y, Resnik P (2008) Soft syntactic constraints for hierarchical phrased-based translation. In: Proceedings of the 44th annual meeting of the Association for Computational Linguistics, Columbus, Ohio, USA, pp 1003–1011

Marton Y, Habash N, Rambow O (2010) Improving Arabic dependency parsing with lexical and inflectional morphological features. In: Proceedings of the 11th Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL) Workshop on statistical parsing of morphologically rich languages (SPMRL), Los Angeles, California

Nivre J (2003) An efficient algorithm for projective dependency parsing. In: Proceedings of the 8th international conference on parsing technologies (IWPT), Nancy, France, pp 149–160

Nivre J (2008) Algorithms for deterministic incremental dependency parsing. Comput Linguist 34(4):513–553

Nivre J, Hall J, Nilsson J (2006) MaltParser: a data-driven parser-generator for dependency parsing. In: Proceedings of the fifth international conference on language resources and evaluation (LREC2006), Genoa, Italy, pp 2216–2219

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–52

Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania pp 311–318

Roth R, Rambow O, Habash N, Diab M, Rudin C (2008) Arabic morphological tagging, diacritization, and lemmatization using lexeme models and feature ranking. In: Proceedings of ACL-08: HLT, short papers, Columbus, Ohio, pp 117–120.

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the conference of the Association for Machine Translation in the Americas, Association for Machine Translation in the Americas, Boston, MA, pp 223–231

Vogel S, Ney H, Tillman C (1996) HMM-based word alignment in statistical machine translation. In: Proceedings of the 16th international conference on computational linguistics, Copenhagen, Denmark, pp 836–841

Wang C, Collins M, Koehn P (2007) Chinese syntactic reordering for statistical machine translation. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, Czech Republic, pp 737–745

Xia F, McCord M (2004) Improving a statistical mt system with automatically learned rewrite patterns. In: Proceedings of the 20th international conference on computational linguistics, Geneva, Switzerland, pp 508–514

Zhang Y, Zens R, Ney H (2007) Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In: Proceedings of the workshop on syntax and structure in statistical translation, Association for Computational Linguistics, Rochester, New York, pp 1–8

Zitouni I, Sorensen JS, Sarikaya R (2006) Maximum entropy based restoration of Arabic diacritics. In: Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics, Sydney, Australia, pp 577–584