TALN 2010, Montréal, 19–23 juillet 2010

Reordering Matrix Post-verbal Subjects for Arabic-to-English SMT

Marine Carpuat Yuval Marton Nizar Habash
Columbia University
Center for Computational Learning Systems
475 Riverside Drive, New York, NY 10115
{marine,ymarton,habash}@ccls.columbia.edu

Résumé. Distinguer les constructions verbe-sujet (VS) des propositions principales ("matrice") et subordonnées ("non-matrice") améliore notre nouveau modèle de réordonnancement pour l'alignement des mots en Traduction Automatique Statistique (TAS) arabe-anglais (Carpuat *et al.*, 2010). D'une part, la majorité des constructions verbe-sujet (VS) dans les propositions principales doivent être réordonnancées en anglais, alors que l'ordre du verbe et du sujet est préservé dans la moitié des cas de constructions VS subordonnées. D'autre part, nous constatons que notre analyseur syntaxique parvient à mieux identifier les constructions VS des propositions principales. Ces observations nous amènent à limiter le réordonnancement des constructions VS à celles des propositions principales lors de l'alignement des mots. Cette technique améliore substantiellement la performance d'un système de TAS conventionnel, et d'un système qui réordonnance toutes les constructions VS. L'amélioration des mesures BLEU et TER obtenue par simple réordonnancement représente presque la moitié de l'amélioration obtenue lorsque le modèle d'alignement des mots est entraîné sur un corpus parallèle d'une taille cinq fois supérieure.

Abstract. We improve our recently proposed technique for integrating Arabic verb-subject constructions in SMT word alignment (Carpuat *et al.*, 2010) by distinguishing between matrix (or main clause) and non-matrix Arabic verb-subject constructions. In gold translations, most matrix VS (main clause verb-subject) constructions are translated in inverted SV order, while non-matrix (subordinate clause) VS constructions are inverted in only half the cases. In addition, while detecting verbs and their subjects is a hard task, our syntactic parser detects VS constructions better in matrix than in non-matrix clauses. As a result, reordering only matrix VS for word alignment consistently improves translation quality over a phrase-based SMT baseline, and over reordering all VS constructions, in both medium- and large-scale settings. In fact, the improvements obtained by reordering matrix VS on the medium-scale setting remarkably represent 44% of the gain in BLEU and 51% of the gain in TER obtained with a word alignment training bitext that is 5 times larger.

Mots-clés : Analyse morpho-syntaxique de l'arabe, Traduction automatique statistique, VS, VSO.

Keywords: Arabic syntactic parsing, Statistical machine translation, VS, VSO.

1 Introduction

Translating Arabic verb subjects into English is challenging for current feature-poor Statistical Machine Translation (SMT) models : Arabic subjects can occur in pre-verbal (SV), post-verbal (VS) or pro-dropped constructions ; gender and number agreement rules differ in SV and VS orders ; and recursive possessive constructions introduce long distance dependencies. This suggests that translating Arabic subjects to English requires complex long-distance reordering. However, standard SMT distance-based reordering models used for word alignment (Och & Ney, 2003) and decoding (Koehn *et al.*, 2007) are not well suited to this task. Most syntax-aware phrase-based models do not capture verb subject information either, as they typically rely on phrase-structure representations (Marton & Resnik (2008) *inter alia*). As a result, subject translations are often garbled, while verbs are frequently incorrectly translated or even dropped.

We have recently shown that reordering VS constructions for word alignment improves Arabic-English translation (Carpuat *et al.*, 2010). However, unlike in previous syntactic reordering approaches, subjects are moved back to the original VS word order before phrase-extraction and decoding. This strategy successfully leverages subject span information, while acknowledging the poor quality of automatic VS detection. 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 the experiments described in (Carpuat *et al.*, 2010), we obtain improvements in translation quality despite using an overly simplistic reordering rule : all VS subjects are reordered, while analysis shows that almost 30% of Arabic VS constructions are translated in the same order in English. In a follow-up analysis, we manually inspected the data and found that many monotone VS occur in subordinate clauses. This suggested that distinguishing between matrix (main) vs. non-matrix (subordinate) subjects might provide additional insights.

In this paper, we show that limiting reordering to matrix VS subjects further improves SMT 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 than reordering all VS.

2 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 verb-subject 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',

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).

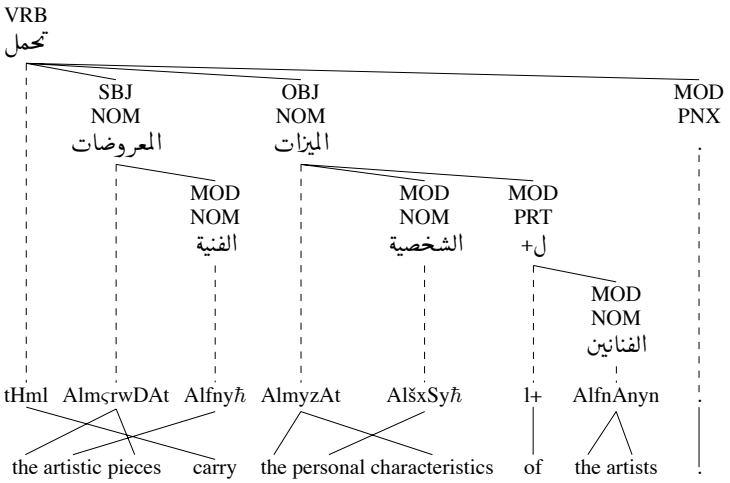

FIGURE 1 – A pair of word-aligned Arabic and English sentences. The Arabic syntactic dependency representation is in CATiB style annotation.

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

[V Aς ln] [SBJ Almnsq Alς Am l+ mšrwς Alskħ AlHdyd byn dwl mjls Altς Awn Alxlyjy] *[OBJ An ...]*
[SBJ The general coordinator of the railroad project among the countries of the Gulf Cooperation Council]
[V announced] *[OBJ that ...]*

FIGURE 2 – An example of long distance reordering of Arabic VS order into English SV order

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 & Sadat, 2006). In this paper 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 Figure 1). We tokenize our data using the Morphological Analysis and Disambiguation for Arabic (MADA) toolkit (Habash & Rambow, 2005), for both parser training purposes and SMT (word alignment, phrase extraction, and decoding).

Second, the subject in Arabic verb / subject 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 Figure 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, or a subject of another verb (e.g., in a subordinate clause). 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 noun phrases (and other potential subjects) but also assign them the correct relation to the correct verb in the sentence.

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

	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

3 Subjects of matrix and non-matrix verbs are reordered differently

We previously reported that, while almost all Arabic SVs are translated in a monotone order in English, the picture is more complex for VS constructions (Carpuat *et al.*, 2010) : the majority of Arabic VS are reordered into English, but 27% are translated in a monotone order. In this paper, we show that reordering patterns of Arabic VS constructions into English are surprisingly different for matrix and non-matrix verbs.

We study the reordering patterns of SV and VS constructions in the manually word-aligned parallel Arabic-English Treebank (LDC2009E82). Given the gold Arabic trees and the gold Arabic-English word alignments, we can determine the gold reorderings for SV and VS constructions. We extract verb and subject representations from the gold constituent parses by deterministic conversion to the simplified dependency structure of the Columbia Arabic Treebank (CATiB) (Habash & Roth, 2009). 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 : Interestingly, VS in matrix clauses are reordered much more frequently (81%) than non-matrix VS (48%). In contrast, both matrix and non-matrix SV almost always translate into a monotone order in English. Manual inspection reveals that the monotone VS translations are mostly explained by changes to passive voice or to non verbal constructions in the English translation.

4 Arabic VS constructions are hard to identify

Before turning to translation, we need to tackle the prerequisite task of identifying Arabic post-verbal subjects, their spans, and the verbs they attach to (and potentially reorder with). Most statistical syntactic parsers that are used in SMT are constituency parsers (Bikel, 2004; Manning & Schuetze, 1999), and do not typically mark subject relations explicitly. In contrast, and as in (Carpuat *et al.*, 2010), we employ a dependency parser – MaltParser with the Nivre "eager" algorithm (Nivre, 2003; Nivre *et al.*, 2006) – 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). Inspired by the Columbia Arabic Treebank (CATiB) (Habash & 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. Further discussion and subsequent work on

the extended CATiB POS tag set can be found in (Marton *et al.*, 2010).

Evaluated on the PATB part 3v3.1 dev set, our parsing model achieves an overall labeled attachment score of 79.25%, using MADA predicted (non-gold) POS tags. However, in this paper, 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 moves *to*. Hence, we argue that combined detection statistics of verbs and their subjects (VATS) constructions are more telling when evaluating parsing quality for reordering. [3] Table 4 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.

Overall, identifying VATS is hard, with 74% F-score. Matrix VATS are much harder to detect – almost 9% absolute lower than non-matrix VATS. Some of the difference is the result of mis-identifying whether the verb is a matrix verb or not. Ignoring matricity, our main target (VS construction) has the lowest performance of all constructions. That said, the VS construction has a much better performance in the harder matrix condition than the non-matrix condition. This is rather different from the other two constructions which fare better in the non-matrix condition.

It should be noted that verb identification (i.e., verb regardless of its subject) is almost perfect (F-score of 99.88% and 100% recall), [4] and that matrix verb identification precision is almost 93%. That said, the low precision of the matrix VNS condition (56.37% in Table 4) suggests that most errors of VS and SV are likely to be errors of unidentified subjects, i.e., the verbs are considered VNS, as opposed to incorrect subject spans).

TABLE 2 – Subject and verb detection Precision, Recall and F scores. 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 percentrage of all VATS in the same matricity condition.

	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

While we intend to work on improving verb-subject detection accuracy, it is also worth exploring whether the current noisy matrix VS detection can still help Arabic-English phrase-based SMT.

3. 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 PATB's. 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.

4. Note that this evaluation starts with gold tokenization.

5 Reordering matrix Arabic VS for SMT word alignment

Based on the analysis of gold reordering patterns and our automatic subject detection tools, we introduce a simple but crucial improvement in the reordering for word alignment method proposed in (Carpuat *et al.*, 2010). This method attempts to make Arabic and English word order closer to each other by reordering Arabic VS constructions into SV during word alignment. However, as we have seen in Section 4, automatic detection of subject boundaries is very noisy, so we adopt a conservative reordering strategy and only reorder Arabic sentences to perform word alignment. Unlike most syntactically motivated reordering models (e.g., Collins *et al.* (2005); Habash (2007)), our system performs phrase-translation extraction and decoding on the original Arabic word order. Reordering Arabic VS attempts to make the bitext easier to explain by the alignment model, and should therefore help generate accurate links between Arabic and English words. 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 correctly detected VS

To sum up, 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 **matrix** VSs into SVs.
3. get word alignment wa' on new sentence pair (a', e) :
4. using mapping from a to $a' = reorder(a)$, get corresponding word alignment $wa = unreorder(wa')$ for the original sentence pair (a, e)

6 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 the GIZA++ (Och & 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. Feature weights 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 & Rambow, 2005). MADA-produced Arabic lemmas are used for word alignment. The dependency parser described in Section 4 is applied to the entire Arabic training data.

7 Matrix VS reordering significantly improves BLEU and TER

On a large test set of more than 4440 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 (Table 3). In addition, restricting reordering to matrix VS also yields better scores than reordering all VS constructions (as in Carpuat *et al.* (2010)). Results per test set

TABLE 3 – Evaluation on all test sets : on the total of 4432 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)

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.

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

BLEU r4n4 (%) / TER (%)					
test set	MT03	MT04	MT05	MT08nw	MT08wb
medium baseline	45.95 / 48.76	44.94 / 46.45	48.05 / 44.99	44.86 / 47.74	32.05 / 58.02
+ matrix VS reordering	46.79 / 47.87	45.28 / 46.15	49.11 / 44.14	45.19 / 47.28	31.98 / 57.34
large baseline	52.30 / 43.33	52.45 / 40.41	54.66 / 39.15	52.60 / 41.81	39.22 / 52.05
+ matrix VS reordering	52.88 / 42.77	52.42 / 40.33	55.29 / 38.74	52.98 / 41.36	40.01 / 52.00

More analysis is needed to better understand how the gains in BLEU and TER relate to the changes in word alignment introduced by matrix VS reordering. Previous work showed that intrinsic evaluation of word alignment quality against manually created references does not correlate well with translation quality (see (Lopez & Resnik, 2006) for an overview.) Since we are primarily interested in the end-goal of improving translation (rather than alignment), we do not compute alignment error rates against manual word alignments. Instead, we use baseline word alignments learned on a much larger training set as a basis for comparison. We argue that these alignments can be used as a valid reference despite being learned automatically, since, unlike gold manual alignments, they do improve translation quality in an end-to-end SMT system.

We therefore build a fourth medium-scale system using word alignment models trained on the large-scale bitext, which is more than 5 times larger than the medium-scale bitext. The SMT system is trained using these improved alignment links for the subset of the large bitext that matches the medium-scale data condition. This 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 a word alignment training bitext that is 5 times larger.

Finally, we compare the word alignment links learned by the different versions of the system on a common sample of about 15k sentence pairs. Table 5 shows that reordering matrix VS yields slightly fewer

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	#links	med baseline	+ all VS	+ matrix VS	large baseline
medium baseline	330255	100%	87.64%	43.28%	66.05%
+ all VS reordering	330255	87.64%	100%	67.75%	58.49%
+ matrix VS reordering	326625	75.51%	67.00%	100%	66.35%

alignment links than both the baseline and the system that reordered all VS. Columns 3-5 show 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. Finally, Column 6 reveals that the matrix VS reordering strategy yields the highest percentage of common links with the improved alignments learned on the large-scale bitext.

8 Related work

To the best of our knowledge, the only other approach to detecting and using Arabic verb-subject (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 (CRF), 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 0.4 BLEU on the Europarl German-English translation task. Xia & 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.

For Arabic-English phrase-based SMT, the impact of syntactic reordering as preprocessing is less clear. Habash (2007) 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 & Habash, 2008).

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.

9 Conclusion

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. 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 in (Carpuat *et al.*, 2010) with a simple but crucial change : instead of reordering all Arabic VS constructions, we limit reordering to matrix VS. This approach remarkably improves the translation quality of strong medium- and large-scale phrase-based SMT systems, despite using noisy matrix VS predictions. The improvements obtained by reordering matrix VS on the medium-scale setting represent 44% of the gain in BLEU and 51% of the gain in TER obtained with a word alignment training bitext that is 5 times larger.

Acknowledgements

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

Références

BIKEL D. M. (2004). Intricacies of Collins' parsing model. *Computational Linguistics*, **30**(4), 479–511.

CARPUAT M., MARTON Y. & HABASH N. (2010). Explorations in subject-verb reordering for arabic-english statistical machine translation. In *Proceedings of ACL-2010*.

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

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)*, p. 531–540.

CREGO J. M. & 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*, p. 53–61, Columbus, Ohio.

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*, p. 44–52.

GREEN S., SATHI C. & MANNING C. D. (2009). NP subject detection in verb-initial Arabic clauses. In *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages (CAASL3)*.

HABASH N. (2007). Syntactic preprocessing for statistical machine translation. In *Proceedings of the Machine Translation Summit (MT-Summit)*, Copenhagen.

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)*, p. 573–580, Ann Arbor, Michigan : Association for Computational Linguistics.

HABASH N. & ROTH R. (2009). CATiB : The columbia arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, p. 221–224, Suntec, Singapore : Association for Computational Linguistics.

HABASH N. & SADAT F. (2006). Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers*, p. 49–52, New York City, USA.

HABASH N., SOUDI A. & BUCKWALTER T. (2007). On Arabic Transliteration. In A. VAN DEN BOSCH & A. SOUDI, Eds., *Arabic Computational Morphology : Knowledge-based and Empirical Methods*. Springer.

HERMJAKOB U. (2009). Improved word alignment with statistics and linguistic heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, p. 229–237, Singapore : Association for Computational Linguistics.

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)*, p. 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 : SIGDAT.

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.

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*, p. 90–99, Cambridge, MA.

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*, p. 102–109, Cairo, Egypt.

MANNING C. D. & SCHUETZE H. (1999). Foundations of statistical natural language processing. p. 304. Cambridge, MA : MIT Press.

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.

MARTON Y. & RESNIK P. (2008). Soft syntactic constraints for hierarchical phrased-based translation. In *Proceedings of ACL-2008*, p. 1003–1011, Columbus, Ohio, USA.

NIVRE J. (2003). An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Conference on Parsing Technologies (IWPT)*, p. 149–160, Nancy, France.

NIVRE J., HALL J. & NILSSON J. (2006). MaltParser : A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

OCH F. J. & NEY H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, **29**(1), 19–52.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

SNOVER M., DORR B., SCHWARTZ R., MICCIULLA L. & MAKHOUL J. (2006). A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, p. 223–231, Boston, MA : Association for Machine Translation in the Americas.

XIA F. & MCCORD M. (2004). Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING 2004*, p. 508–514, Geneva, Switzerland.

ZHANG Y., ZENS R. & NEY H. (2007). Chunk-level reordering of source language sentences with automatically learned rules for statistical machine translation. In *Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting*, Rochester, NY.

ZITOUNI I., SORENSEN J. S. & 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*, p. 577–584, Sydney, Australia.