Improving Arabic-to-English Statistical Machine Translation by Reordering Post-verbal Subjects for Alignment

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

Abstract

We study the challenges raised by Arabic verb and subject detection and reordering in Statistical Machine Translation (SMT). We show that post-verbal subject (VS) constructions are hard to translate because they have highly ambiguous reordering patterns when translated to English. In addition, implementing reordering is difficult because the boundaries of VS constructions are hard to detect accurately, even with a state-of-the-art Arabic dependency parser. We therefore propose to reorder VS constructions into SV order for SMT word alignment only. This strategy significantly improves BLEU and TER scores, even on a strong large-scale baseline and despite noisy parses.

1 Introduction

Modern Standard Arabic (MSA) is a morpho-syntactically complex language, with different phenomena from English, a fact that raises many interesting issues for natural language processing and Arabic-to-English statistical machine translation (SMT). While comprehensive Arabic preprocessing schemes have been widely adopted for handling Arabic morphology in SMT (e.g., Sadat and Habash (2006), Zollmann et al. (2006), Lee (2004)), syntactic issues have not received as much attention by comparison (Green et al. (2009), Crego and Habash (2008), Habash (2007)). Arabic verbal constructions are particularly challenging since subjects can occur in pre-verbal (SV), post-verbal (VS) or pro-dropped ("null subject") constructions. As a result, training data for learning verbal construction translations is split between the different constructions and their patterns; and complex reordering schemas are needed in order to translate them into primarily pre-verbal subject languages (SVO) such as English.

These issues are particularly problematic in phrase-based SMT (Koehn et al., 2003). Standard phrase-based SMT systems memorize phrasal translation of verb and subject constructions as observed in the training bitext. They do not capture any generalizations between occurrences in VS and SV orders, even for the same verbs. In addition, their distance-based reordering models are not well suited to handling complex reordering operations which can include long distance dependencies, and may vary by context. Despite these limitations, phrase-based SMT systems have achieved competitive results in Arabic-to-English benchmark evaluations.[1] However, error analysis shows that verbs are still often dropped or incorrectly translated, and subjects are split or garbled in translation. This suggests that better syntactic modeling should further improve SMT.

We attempt to get a better understanding of translation patterns for Arabic verb constructions, 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 subject translation. In order to detect VS constructions, we use our state-of-the-art Arabic dependency parser, which is essentially the CATIBEX baseline in our subsequent parsing work in Marton et al. (2010), and is further described there. We show that VS subjects and their exact boundaries are hard to identify accurately. Given the 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 de-

[1] http://www.itl.nist.gov/iad/mig/tests/mt/2009/ResultsRelease/currentArabic.html

Proceedings of the ACL 2010 Conference Short Papers, pages 178–183,
Uppsala, Sweden, 11-16 July 2010.

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	%
SV	monotone	2588	**98.2**
SV	inverted	15	0.5
SV	overlap	35	1.3
SV	total	2638	100
VS	monotone	1700	27.3
VS	inverted	4033	**64.7**
VS	overlap	502	8.0
VS	total	6235	100

pendency parses as cohesion constraints in decoding (e.g., 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.

In this work, we focus on VS constructions only, and propose a new strategy in order to benefit from their noisy detection: for the word alignment stage only, we reorder phrases detected as VS constructions into an SV order. Then, for phrase extraction, weight optimization and decoding, we use the original (non-reordered) text. This approach significantly improves both BLEU and TER on top of strong medium and large-scale phrase-based SMT baselines.

2 VS reordering in gold Arabic-English translation

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 Section 3). 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 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 verb subjects occur more frequently in VS order (70%) than in SV order (30%). These numbers do not include pro-dropped ("null subject") constructions.

3 Arabic VS construction detection

Even if the SMT system had perfect knowledge of VS reordering, it has to accurately detect VS constructions and their spans in order to apply the reordering correctly. For that purpose, we use our state-of-ther-art parsing model, which is essentially the CATIBEX baseline model in Marton et al. (2010), and whose details we summarize next. We train a syntactic dependency parser, MaltParser v1.3 with the Nivre "eager" algorithm (Nivre, 2003; Nivre et al., 2006; Nivre, 2008) on the training portion of the Penn Arabic Treebank part 3 v3.1, hereafter PATB3 (Maamouri et al., 2008; Maamouri et al., 2009). The training / development split is the same as in Zitouni et al. (2006). We convert the PATB3 representation into the succinct CATiB format, with 8 dependency relations and 6 POS tags, which we then extend to a set of 44 tags using regular expressions of the basic POS and the normalized surface word form, similarly to Marton et al. (2010), following Habash and Roth (2009). We normalize Alif Maqsura to Ya, and Hamzated Alifs to bare Alif, as is commonly done in Arabic SMT.

For analysis purposes, we evaluate our subject and verb detection on the development part of PATB3 using gold POS tags. There are various ways to go about it. We argue that combined detection statistics of constructions of verbs and their subjects (VATS), for which we achieve an F-score of 74%, are more telling for the task at hand.[2]

[2]We divert from the CATiB representation in that a non-matrix subject of a pseudo verb (*An* and her sisters) 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.

These scores take into account the spans of both the subject and the specific verb it belongs to, and potentially reorder with. We also provide statistics of VS detection separately (F-score 63%), since we only handle VS here. This low score can be explained by the difficulty in detecting the post-verbal subject's end boundary, and the correct verb the subject belongs to. The SV construction scores are higher, presumably since the pre-verbal subject's end is bounded by the verb it belongs to. See Table 2.

Although not directly comparable, our VS scores are similar to those of Green et al. (2009). Their VS detection technique with conditional random fields (CRF) is different from ours in bypassing full syntactic parsing, and in only detecting maximal (non-nested) subjects of verb-initial clauses. Additionally, they use a different training / test split of the PATB data (parts 1, 2 and 3). They report 65.9% precision and 61.3% F-score. Note that a closer score comparison should take into account their reported verb detection accuracy of 98.1%.

Table 2: Precision, Recall and F-scores for constructions of Arabic verbs and their subjects, evaluated on our development part of PATB3.

construction	P	R	F
VATS (verbs & their subj.)	73.84	74.37	74.11
VS	66.62	59.41	62.81
SV	86.75	61.07	71.68
VNS (verbs w/ null subj.)	76.32	92.04	83.45
verbal subj. exc. null subj.	72.46	60.18	65.75
verbal subj. inc. null subj.	73.97	74.50	74.23
verbs with non-null subj.	91.94	76.17	83.31
SV or VS	72.19	59.95	65.50

4 Reordering Arabic VS for SMT word alignment

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 attempts 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. Preliminary experiments on an earlier version of the large-scale SMT system described in Section 6 showed that forcing reordering of all VS constructions at training and test time does not have a consistent impact on translation quality: for instance, on the NIST MT08-NW test set, TER slightly improved from 44.34 to 44.03, while BLEU score decreased from 49.21 to 49.09.

Limiting reordering to alignment allows the system to be more robust and recover from incorrect changes introduced either by incorrect VS detection, or by incorrect reordering of a correctly detected VS. 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)

5 Experiment 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:

- **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.
- **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.

Except from this difference in training data, the two systems are identical. They use a standard phrase-based architecture. The parallel corpus is word-aligned using the 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. We apply statistical significance tests to prune unreliable phrase-pairs

and score remaining phrase-table entries (Chen et al., 2009). 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.

For all systems, the English data is tokenized using simple punctuation-based rules. The Arabic side is segmented according to the Arabic Treebank (PATB3) tokenization scheme (Maamouri et al., 2009) using the MADA+TOKAN morphological analyzer and tokenizer (Habash and Rambow, 2005). MADA-produced Arabic lemmas are used for word alignment.

6 Results

We evaluate translation quality using both BLEU (Papineni et al., 2002) and TER (Snover et al., 2006) scores on three standard evaluation test sets from the NIST evaluations, which yield more than 4400 test sentences with 4 reference translations. On this large data set, our VS reordering method remarkably yields statistically significant improvements in BLEU and TER on the medium and large SMT systems at the 99% confidence level (Table 3).

Results per test set are reported in Table 4. TER scores are improved in all 10 test configurations, and BLEU scores are improved in 8 out of the 10 configurations. Results on the MT08 test set show that improvements are obtained both on newswire and on web text as measured by TER (but not BLEU score on the web section.) It is worth noting that consistent improvements are obtained even on the large-scale system, and that both baselines are full-fledged systems, which include lexicalized reordering and large 5-gram language models.

Analysis shows that our VS reordering technique improves word alignment coverage (yielding 48k and 330k additional links on the medium and large scale systems respectively). This results in larger phrase-tables which improve translation quality.

7 Related work

To the best of our knowledge, the only other approach to detecting and using Arabic verb-subject constructions for SMT is that of Green et al. (2009) (see Section 3), which failed to improve Arabic-English SMT. In contrast with our reordering approach, they integrate subject span information as a log-linear model feature which encourages a phrase-based SMT decoder to use phrasal translations that do not break subject boundaries.

Table 3: Evaluation on all test sets: on the total of 4432 test sentences, improvements are statistically significant at the 99% level using bootstrap resampling (Koehn, 2004)

system	BLEU r4n4 (%)	TER (%)
medium baseline	44.35	48.34
+ VS reordering	44.65 (+0.30)	47.78 (-0.56)
large baseline	51.45	42.45
+ VS reordering	51.70 (+0.25)	42.21 (-0.24)

Syntactically motivated reordering for phrase-based SMT has been more successful on language pairs other 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 improve German-English translation by 0.4 BLEU on the Europarl task. 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.

For Arabic-English phrase-based SMT, the impact of syntactic reordering as preprocessing is less clear. Habash (2007) proposes to learn syntactic reordering rules targeting Arabic-English word order differences and integrates them as deterministic preprocessing. He reports improvements in BLEU compared to phrase-based SMT limited to monotonic decoding, but these improvements do not hold with distortion. Instead of applying reordering rules deterministically, Crego and Habash (2008) use a lattice input to represent alternate word orders which improves a ngram-based SMT system. But they do not model VS constructions explicitly.

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

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 (%)					
test set	MT03	MT04	MT05	MT08nw	MT08wb
medium baseline	45.95	44.94	48.05	44.86	32.05
+ VS reordering	46.33 (+0.38)	45.03 (+0.09)	48.69 (+0.64)	45.06 (+0.20)	31.96 (-0.09)
large baseline	52.3	52.45	54.66	52.60	39.22
+ VS reordering	52.63 (+0.33)	52.34 (-0.11)	55.29 (+0.63)	52.85 (+0.25)	39.87 (+0.65)
TER (%)					
test set	MT03	MT04	MT05	MT08nw	MT08wb
medium baseline	48.77	46.45	45.00	47.74	58.02
+ VS reordering	48.31 (-0.46)	46.10 (-0.35)	44.29 (-0.71)	47.11 (-0.63)	57.30 (-0.72)
large baseline	43.33	40.42	39.15	41.81	52.05
+ VS reordering	42.95 (-0.38)	40.40 (-0.02)	38.75 (-0.40)	41.51 (-0.30)	51.86 (-0.19)

links if they degrade their syntax-based translation system. Departing from commonly-used alignment models, 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.

8 Conclusion and future work

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

In subsequent work, we show that matrix (main-clause) VS constructions are reordered much more frequently than non-matrix VS, and that limiting reordering to matrix VS constructions for word alignment further improves translation quality (Carpuat et al., 2010). In the future, we plan to improve robustness to parsing errors by using not just one, but multiple subject boundary hypotheses. We will also investigate the integration of VS reordering in SMT decoding.

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 authors and do not necessarily reflect the views of DARPA.

References

Nguyen Bach, Stephan Vogel, and Colin Cherry. 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*, pages 1–4.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Reordering matrix post-verbal subjects for arabic-to-english smt. In *Proceedings of the Conference Traitement Automatique des Langues Naturelles (TALN)*.

Boxing Chen, George Foster, and Roland Kuhn. 2009. Phrase translation model enhanced with association based features. In *Proceedings of MT-Summit XII*, Ottawa, Ontario, September.

Colin Cherry. 2008. Cohesive phrase-based decoding for statistical machine translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 72–80, Columbus, Ohio, June.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause restructuring for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 531–540, Ann Arbor, MI, June.

Josep M. Crego and Nizar Habash. 2008. Using shallow syntax information to improve word alignment and reordering for SMT. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 53–61, June.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 44–52.

Spence Green, Conal Sathi, and Christopher D. Manning. 2009. NP subject detection in verb-initial Arabic clauses.

In *Proceedings of the Third Workshop on Computational Approaches to Arabic Script-based Languages (CAASL3)*.

Nizar Habash and Owen Rambow. 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)*, pages 573–580, Ann Arbor, Michigan, June.

Nizar Habash and Ryan Roth. 2009. CATiB: The Columbia Arabic treebank. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 221–224, Suntec, Singapore, August. Association for Computational Linguistics.

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

Ulf Hermjakob. 2009. Improved word alignment with statistics and linguistic heuristics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 229–237, Singapore, August.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT/NAACL-2003*, Edmonton, Canada, May.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, Prague, Czech Republic, June.

Philipp Koehn. 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, July.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 57–60, Boston, MA.

Mohamed Maamouri, Ann Bies, and Seth Kulick. 2008. Enhancing the arabic treebank: a collaborative effort toward new annotation guidelines. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, and Basma Bouziri. 2009. The penn arabic treebank part 3 version 3.1. Linguistic Data Consortium LDC2008E22.

Yuval Marton, Nizar Habash, and Owen Rambow. 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.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A Data-Driven Parser-Generator for Dependency Parsing. In *Proceedings of the Conference on Language Resources and Evaluation (LREC)*.

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

Joakim Nivre. 2008. Algorithms for Deterministic Incremental Dependency Parsing. *Computational Linguistics*, 34(4).

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–52.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Fatiha Sadat and Nizar Habash. 2006. Combination of arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1–8, Morristown, NJ, USA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of AMTA*, pages 223–231, Boston, MA.

Chao Wang, Michael Collins, and Philipp Koehn. 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)*, pages 737–745.

Fei Xia and Michael McCord. 2004. Improving a statistical mt system with automatically learned rewrite patterns. In *Proceedings of COLING 2004*, pages 508–514, Geneva, Switzerland, August.

Yuqi Zhang, Richard Zens, and Hermann Ney. 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, April.

Imed Zitouni, Jeffrey S. Sorensen, and Ruhi Sarikaya. 2006. Maximum Entropy Based Restoration of Arabic Diacritics. In *Proceedings of COLING-ACL, the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics*, pages 577–584, Sydney, Australia.

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. 2006. Bridging the inflection morphology gap for arabic statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 201–204, New York City, USA.