TALN 2010, Montréal, 19–23 juillet 2010

Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation

Ahmed El Kholy Nizar Habash
Center for Computational Learning Systems, Columbia University
{akholy,habash}@ccls.columbia.edu

Résumé. De nombreux travaux en Traduction Automatique Statistique (TAS) pour des langues d'entrée morphologiquement riches montrent que la ségmentation morphologique et la normalisation orthographique améliorent la qualité des traductions en diminuant la sparsité des données. Dans cet article, nous étudions l'impact de ce prétraitement pour la TAS vers une langue de sortie riche morphologiquement, comme l'Arabe. Nous explorons l'espace des schémas de segmentation et des options de normalisation possibles. Nous évaluons seulement la sortie sous une forme désegmentée et enrichie orthographiquement. Nos résultats montrent d'une part que le meilleur schéma pour la ségmentation est celui de la Penn Arabic Treebank. D'autre part, la meilleure procédure de prétraitement consiste à entraîner le système sur des données normalisées orthographiquement, puis à enrichir et désegmenter les traductions en sortie.

Abstract. Much of the work on Statistical Machine Translation (SMT) from morphologically rich languages has shown that morphological tokenization and orthographic normalization help improve SMT quality because of the sparsity reduction they contribute. In this paper, we study the effect of these processes on SMT when translating into a morphologically rich language, namely Arabic. We explore a space of tokenization schemes and normalization options. We only evaluate on detokenized and orthographically correct (enriched) output. Our results show that the best performing tokenization scheme is that of the Penn Arabic Treebank. Additionally, training on orthographically normalized (reduced) text then jointly enriching and detokenizing the output outperforms training on enriched text.

Mots-clés : Langue Arabe, Morphologie, Ségmentation, Désegmentation, La Traduction Automatique Statistique.

Keywords: Arabic Language, Morphology, Tokenization, Detokenization, Statistical Machine Translation.

1 Introduction

Most of the published research on statistical machine translation (SMT) for Arabic focuses on Arabic-to-English. Recently, however, translation into Arabic has been receiving increasing attention (Sarikaya & Deng, 2007; Badr *et al.*, 2008, 2009; Elming & Habash, 2009; El Kholy & Habash, 2010). In this paper, we consider some of the lessons learned from working on Arabic-English MT and their applicability to the English-Arabic direction. In particular we focus on two techniques : morphological tokenization (for short tokenization) and orthographic normalization (for short normalization). The common wisdom in the field of natural language processing (NLP) is that tokenization of Arabic words through decliticization and reductive orthographic normalization is helpful for MT into English because of the sparsity reduction they contribute. We explore a space of tokenization schemes and normalization options and their implications on the quality of English-Arabic MT. Regardless of the preprocessing choices, the Arabic output is detokenized and denormalized. Anything less is comparable to producing all lower cased English or uncliticized and undiacritized French. The focus of this paper is on the value of tokenization and normalization for SMT. For a detailed discussion of detokenization and denormalization for Arabic, see (El Kholy & Habash, 2010).

This paper is organized as follows. Section 2 briefly discusses some related work. Section 3 introduces relevant linguistic issues. Sections 4 and 5 present our approach and experimental results, respectively.

2 Related Work

Much of the work done on studying the effects of morphological preprocessing on SMT quality focuses on translation from morphologically rich languages such as German (Nießen & Ney, 2004), Czech (Goldwater & McClosky, 2005) and Arabic (Habash & Sadat, 2006; Lee, 2004; Zollmann *et al.*, 2006). They show that reducing the sparsity caused by rich morphology through some form of morphological tokenization has a positive impact on the quality of SMT . There is also a growing number of publications that consider translation into morphologically rich languages such as Turkish (Oflazer & Durgar El-Kahlout, 2007) and Arabic (Sarikaya & Deng, 2007; Badr *et al.*, 2008; El Kholy & Habash, 2010). We focus here on efforts that studied the impact of morphological preprocessing on Arabic as a target language. Sarikaya & Deng (2007) use joint morphological-lexical language models to re-rank the output of English-dialectal Arabic MT. Badr *et al.* (2008) report results on the value of morphological tokenization of Arabic during training and describe different techniques for detokenizing Arabic output. The research presented here is most closely related to that of Badr *et al.* (2008). We extend on their contribution in two ways : (a.) we present a comparison of a larger number of tokenization schemes that yielded improved results over theirs ; and (b) we discuss the technical challenges and solutions of producing unnormalized Arabic output whereas they only reported results on normalized Arabic. We will compare to their work as appropriate throughout the paper. Finally, in a previous publication (El Kholy & Habash, 2010) , we presented an extension to Badr *et al.* (2008)'s work on detokenization. We do not discuss the details of this work here ; however, we use our best previously reported detokenization setting for the experiments in this paper (see Section 4.3).

3 Arabic Linguistic Issues

In this section we present relevant aspects of Arabic word orthography and morphology.

Rule Name	Tokenized	Untokenized	Example		
			Tokenized	Untokenized	Gloss
Definite Article	?ل+ال+ل *l+Al+l ?*	+لل *ll+*	مكتب+ال+ل *l+Al+mktb*	للمكتب *llmktb*	'for the office'
			الجنة+ال+ل *l+Al+ljnħ*	للجنة *lljnħ*	'for the committee'
Ta-Marbuta	ة- *-ħ* +pron	ت- *-t* +pron	مكتبة+هم *mktbħ+hm*	مكتبتهم *mktbthm*	'their library'
Alif-Maqsura	ى- *-ý* +pron	ا- *-A* +pron	روى+ه *rwý+h*	رواه *rwAh*	'he watered it'
	exceptionally	ي- *-y* +pron	على+ه *ςlý+h*	عليه *ςlyh*	'on him'
Hamza	ء- *-'* +pron	ئ- *-ŷ* +pron	بهاء+ه *bhA'+h*	بهائه *bhAŷh*	'his glory [gen.]'
	less frequently	ؤ- *-ŵ* +pron	بهاء+ه *bhA'+h*	بهاؤه *bhAŵh*	'his glory [nom.]'
	less frequently	ء- *-'* +pron	بهاء+ه *bhA'+h*	بهاءه *bhA'h*	'his glory [acc.]'

TABLE *1* – Examples of some Arabic morphological adjustment rules

3.1 Arabic Orthography

Certain letters in Arabic script are often spelled inconsistently which leads to an increase in both sparsity (multiple forms of the same word) and ambiguity (same form corresponding to multiple words). In particular, variants of Hamzated Alif, أ Â [1] or إ Ă, are often written without their Hamza (ء *'*) : ا *A* ; and the Alif-Maqsura (or dotless Ya) ى *ý* and the regular dotted Ya ي *y* are often used interchangeably in word final position. This inconsistent variation in raw Arabic text is typically addressed in Arabic NLP through what is called orthographic normalization, a reductive process that converts all Hamzated Alif forms (including Alif Madda آ Ā) to bare Alif and dotless Ya/Alif Maqsura form to dotted Ya. This kind of normalization is referred to as a Reduced normalization (RED). RED normalization is contrasted with Enriched normalization (ENR), which selects the appropriate form of the Alif and Ya in context (El Kholy & Habash, 2010). ENR Arabic is optimally the desired form of Arabic to generate and to evaluate against. Comparing a manually enriched (ENR) version of the Penn Arabic Treebank (PATB) (Maamouri *et al.*, 2004) to its reduced (RED) version, we find that 16.2% of the words are different. However, the raw (naturally unnormalized) version of the PATB is only different in 7.4% of the words. This suggests a major problem in the recall of the correct ENR form in raw text. In internal experiments, we noticed that BLEU-4 (Papineni *et al.*, 2002) scores drop about 10 % absolute when comparing ENR to raw (as opposed to ENR) and about 5 % when comparing RED to raw (as opposed to RED) for the same output. As such we only evaluate results against references with their matching normalization condition (ENR or RED).

3.2 Arabic Morphology

Arabic is a morphologically complex language with a large set of morphological features producing a large number of rich word forms. While the number of (morphologically untokenized) Arabic words in a parallel corpus is 20% less than the number of corresponding English words, the number of unique Arabic word types is over twice the number of unique English word types over the same corpus size.

One aspect of Arabic that contributes to this complexity is its various attachable clitics. We define three degrees of cliticization that are applicable in a strict order to a word base (Habash & Sadat, 2006) :

```
[cnj+ [prt+ [art+ BASE +pron]]]
```

1. All Arabic transliterations are provided in the Habash-Soudi-Buckwalter transliteration scheme (Habash *et al.*, 2007).

At the deepest level, the BASE can have either the definite article (+ال *Al*+ 'the') or a member of the class of pronominal enclitics, +pron, (e.g., +هم *+hm* 'their/them'). Next comes the class of particle proclitics (prt+), e.g., +ل *l*+ 'to/for'. At the shallowest level of attachment we find the conjunction proclitic (cnj+), e.g., +و *w*+ 'and'. The attachment of clitics to word forms is not a simple concatenation process. There are several orthographic and morphological adjustment rules that are applied to the word. Some examples of these rules are presented in Table 1 and an almost complete list of the rules is presented and exemplified in (El Kholy & Habash, 2010).

It is important to make the distinction here between simple word segmentation, which splits off word substrings with no orthographic/morphological adjustments, and morphological tokenization, which does. Although segmentation by itself can have important advantages, it leads to the creation of inconsistent or ambiguous word forms : consider the words مكتبة *mktbħ* 'a library' and مكتبتهم *mktbthm* 'their library'. A simple segmentation of the second word creates the non-word string مكتبت *mktbt* ; however, applying adjustment rules as part of the tokenization generates the same form of the basic word in the two cases. See example of Ta-Marbuta rule in Table 1. In this paper, we do not explore morphological tokenization beyond decliticization (details in the next section).

4 Approach

We would like to study the value of a variety of tokenization schemes and orthographic normalizations on English-Arabic SMT. Yet our goal is to always produce correctly detokenized and orthographically enriched Arabic words. To that end, we consider the following variants.

4.1 Orthographic Normalization

Throughout our experiments, we consider two kinds of orthographic normalization schemes, enriched Arabic (ENR) and reduced Arabic (RED) ; but we always target enriched Arabic ENR. For RED systems, we jointly detokenize and enrich the output as explained later in this section.

For comparison reasons, we report results in both RED and ENR Arabic forms. We only compare in the matching form, i.e., RED hypothesis to RED reference and ENR hypothesis to ENR reference.

4.2 Morphological Tokenization

We consider five tokenization schemes discussed in the literature, in addition to a baseline no-tokenization scheme (D0). The D1, D2, TB and D3 schemes were first presented by Habash & Sadat (2006) and the S2 scheme was presented by Badr *et al.* (2008). The S1 scheme used by Badr *et al.* (2008) is the same as Habash & Sadat (2006)'s D3 scheme. TB is the PATB tokenization scheme. We use the Morphological Analysis and Disambiguation for Arabic (MADA) toolkit (Habash & Rambow, 2005) to produce the various tokenization schemes. Figure 1 illustrates the different tokenization schemes with an example. Table 2 presents definitions and various relevant statistics for each tokenization scheme. The schemes differ widely in terms of the increase of number of tokens and the corresponding type count reduction. The more verbose schemes, i.e., schemes with more splitting and higher number of word tokens, have a lower number of token types, which leads to lower out-of-vocabulary (OOV) rates and lower perplexity ; however, they are also harder to predict correctly. The increase in the number of tokens has consequences on word alignment, translation models and language models (LM). We control for these effects in our experiments in Section 5.

Arabic	وسينهى الرئيس جولته بزيارة الى تركيا.						
	wsynhý	Alrŷys	jwlth	bzyArħ	Alý	trkyA	.
Gloss	and will finish	the president	tour his	with visit	to	Turkey	.
English	The president will finish his tour with a visit to Turkey.						
Scheme							
D0	wsynhy	Alrŷys	jwlth	bzyArħ	Ălý	trkyA	.
D1	w+ synhy	Alrŷys	jwlth	bzyArħ	Ălý	trkyA	.
D2	w+ s+ ynhy	Alrŷys	jwlth	b+ zyArħ	Ălý	trkyA	.
TB	w+ s+ ynhy	Alrŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
S2	w+s+ ynhy	Al+ rŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
D3	w+ s+ ynhy	Al+ rŷys	jwlħ +h	b+ zyArħ	Ălý	trkyA	.
LEM	Ânhý	rŷys	jwlħ	zyArħ	Ălý	trkyA	.

FIGURE *1* – A sentence in the various tokenization schemes discussed in Section 4. All tokenizations are in ENR normalization, but the original Arabic is in raw normalization. **LEM** is the lemma form of each word (discussed in Section 5.2).

	Definition	**Change Relative to D0**			**OOV**		**Perplexity**		**Prediction Error Rate**		
		Token#	ENR Type#	RED Type#	ENR	RED	ENR	RED	ENR	RED	SEG
D0	word				2.22	2.17	412.3	410.6	0.62	0.09	0.00
D1	cnj+ word	+7.2	-17.6	-17.8	1.91	1.89	259.3	258.2	0.76	0.23	0.14
D2	cnj+ prt+ word	+13.3	-32.3	-32.6	1.50	1.50	185.5	184.7	0.89	0.37	0.25
TB	cnj+ prt+ word +pron	+17.9	-43.9	-44.2	1.22	1.22	142.2	141.5	1.07	0.57	0.42
S2	cnj+prt+art word +pron	+40.6	-53.0	-53.3	0.91	0.91	69.3	69.0	1.20	0.73	0.60
D3	cnj+ prt+ art+ word +pron	+44.2	-53.0	-53.3	0.90	0.90	61.9	61.7	1.20	0.73	0.60

TABLE *2* – A comparison of the different tokenization schemes studied in this paper : tokenization scheme definition ; the relative change from no-tokenization (D0) in tokens (Token#) and enriched and reduced word types (ENR Type# and RED Type#, respectively) ; out-of-vocabulary (OOV) rate ; perplexity ; MADA's prediction error rate for enriched tokens, reduced tokens and just segmentation (SEG). OOV rates and perplexity values are measured against the NIST MT04 test set while prediction error rates are measured against a Penn Arabic Treebank devset.

4.3 Detokenization and Orthographic Enrichment

All of the systems trained with tokenizations other than D0 require detokenization into correct Arabic word forms. After exploring a set of techniques discussed in (El Kholy & Habash, 2010), we use their best detokenization technique labeled T+R+LM. The technique crucially utilizes a lookup table (T) mapping tokenized forms to detokenized forms. The table is based on pairs of tokenized and detokenized words from our LM data which had been processed by MADA (Habash & Rambow, 2005). In run time, alternatives are given different conditional probabilities, $P(detokenized|tokenized)$, derived from the tables. Tokenized words absent from the tables are detokenized using deterministic rules (R) such as those exemplified in Table 1 as a back off strategy. We use a 5-gram untokenized-form LM and the `disambig` utility in the SRILM toolkit (Stolcke, 2002) to decide among different alternatives. To jointly detokenize and enrich, special tables mapping tokenized RED words to detokenized ENR words are used with ENR language models (ENR-LM).

We compare the performance of the detokenization technique (T+R+LM) for different tokenization schemes and normalization conditions. The results are measured against the Arabic side of the NIST MTEval MT04+MT05 test sets, which together have 2,409 sentences comprising 64,554 words. We report the results in Table 3 in terms of word-level detokenization error rate (WER) and sentence-level detokenization error rate (SER). SER is the percentage of sentences with at least one incorrectly detokenized word.

Overall, the more verbose schemes are harder to detokenize. There are some exceptions, however. D1 and D2 results are exactly the same when the detokenization is made under the same normalization condition. In the case of detokenizing and enriching jointly, D2 shows a slight improvement. This could be explained by the fact that tokenization is helpful in disambiguating some cases in interaction with the LM. However, with more splitting, the task becomes harder and the benefits are diminished. This could be especially noticed in the jump of WER and SER between D2 and TB. This jump is tied to the harder detokenization decisions associated with pronominal enclitics (+pron). All the results of the joint detokenization and enrichment are very close because they all depend on the same LM. S2 and D3 only vary in clustering proclitics and are exactly the same otherwise as far as detokenization is concerned.

Scheme	ENR		RED			
			RED		ENR-LM	
	WER	SER	WER	SER	WER	SER
D1	0.003	0.083	0.003	0.083	0.242	5.853
D2	0.006	0.166	0.006	0.166	0.235	5.687
TB	0.020	0.540	0.023	0.623	0.243	5.853
S2/D3	0.022	0.581	0.025	0.664	0.245	5.895

TABLE *3* – Detokenization and enrichment results for different tokenization schemes in terms of word-level detokenization error rate (WER) and sentence-level detokenization error rate (SER).

5 Experiments

5.1 Experimental Data

All of the training data we use is available from the Linguistic Data Consortium (LDC). [2] We use an English-Arabic parallel corpus of about 142K sentences and 4.4 million words for translation model training data. The parallel text includes Arabic News (LDC2004T17), eTIRR (LDC2004E72), English translation of Arabic Treebank (LDC2005E46), and Ummah (LDC2004T18). Word alignment is done using GIZA++ (Och & Ney, 2003). For language modeling, we use 200M words from the Arabic Gigaword Corpus (LDC2007T40) together with the Arabic side of our training data. Twelve LMs were built for all combinations of normalization and tokenization schemes. We used 5-grams for all LMs unlike Badr *et al.* (2008), who used different n-grams sizes for tokenized and untokenized variants. All LMs are implemented using the SRILM toolkit (Stolcke, 2002).

MADA is used to preprocess the Arabic text for translation modeling and language modeling to produce enriched forms and tokenizations. It can predict the correct enriched form of Arabic words at 99.4%. [3] English preprocessing simply includes down-casing, separating punctuation and splitting off "'s".

2. http ://www.ldc.upenn.edu
3. Statistics are measured on a devset from the Penn Arabic Treebank (Maamouri *et al.*, 2004).

Align	Translation Model	Post Process	Reference Matching
Lemma	ENR	ENR	**25.25**
		RED	**25.29**
	RED	ENR-LM	24.89
		RED	24.96
Surface	ENR	ENR	**24.90**
		RED	**24.97**
	RED	ENR-LM	24.58
		RED	24.68

TABLE *4* – Baseline SMT experiments with D0 tokenization. All results are in BLEU.

Due to the fact that the number of tokens per sentence changes from one tokenization scheme to another, GIZA++'s initial sentence-length filters will drop more sentences from the more verbose schemes. The percentage of sentences dropped due to the filtration process can be up to 2.3% in D3 (versus D0) for a generic cut off of 100 tokens per sentence in Arabic. It may seem like a small percentage ; but since all dropped sentences are very long, this leads to D0 having access to 6.6 % extra words in training over D3. To control for this issue, we filter the training data so that all experiments are done on the same sentences. We use the D3 tokenization scheme as a reference and set the cutoff at 100 D3 tokens.

All experiments are conducted using the Moses phrase-based SMT system (Koehn *et al.*, 2007). The decoding weight optimization was done using a set of 300 sentences from the 2004 NIST MT evaluation test set (MT04). The tuning is based on tokenized Arabic without detokenization. We use a maximum phrase length of size 8 for all tokenizations. We report results on the 2005 NIST MT evaluation set (MT05). These test sets were created for Arabic-English MT and have 4 English references. We use only one Arabic reference in reverse direction for both tuning and testing. We evaluate using BLEU-4 (Papineni *et al.*, 2002) although we are aware of its caveats (Callison-Burch *et al.*, 2006).

5.2 Baseline System

For our baseline system, using D0 tokenization, we compare the value of using lemmas for automatic word alignment as opposed to word surface forms (ENR or RED). In both cases, the phrase tables are built using the surface forms. We compare different combinations of settings for translation models and post-processing. For translation models, we either train on ENR or RED text. As for post-processing, we either keep the output as is, reduce it or enrich it. The enrichment is done using a variant of the ENR-LM technique discussed in Section 4.3. In this experiment set, we did not use the D3-based length filtering described above. The D3-based length filtering is used in the next section. The results in Table 4 show that lemma-based alignment consistently yields superior results to surface-based alignment for the same translation model and post-processing conditions. The rest of the experiments in this paper will all use lemma-based alignment in the following manner : when aligning a verbose tokenization, the lemma form will be used instead of the base word and the separated clitics will not be modified. Table 4 also shows that ENR training is better than RED training ; however, since automatic enrichment error increases with tokenization verbosity (see Table 2, column 10), it is not clear which normalization settings is best to use with verbose schemes. We explore these combinations next.

System	ENR		RED	
Evaluation	ENR	RED	ENR-LM	RED
D0	24.63	24.67	24.66	24.71
D1	25.92	25.99	26.06	26.12
D2	26.41	26.49	26.06	26.15
TB	**26.46**	**26.51**	**26.73**	**26.80**
S2	25.71	25.76	26.11	26.19
D3	25.68	25.75	25.03	25.10

TABLE *5* – Comparing different tokenizations schemes on 4M data sets

5.3 Tokenization Experiments

We compare the performance of the different tokenization schemes and normalization conditions. The results are presented in Table 5. The best performer across all conditions is the TB scheme. The previously reported best performer was S2 (Badr *et al.*, 2008), which was only compared against D0 and D3 tokenizations. Our results are consistent with Badr *et al.* (2008)'s results regarding D0 and D3. However, our TB result outperforms S2. The differences between TB and all other conditions are statistically significant above the 95% level. Statistical significance is computed using paired bootstrap resampling (Koehn, 2004). Training over RED Arabic then enriching its output sometimes yields better results than training on ENR directly which is the case with the TB tokenization scheme. However, sometimes the opposite is true as demonstrated in the D3 results. This is likely due to a tradeoff between the quality of translation and the quality of detokenization.

Contrasting the results in Table 5 for D0 with the first half (Lemma) in Table 4 highlights the effect of length-based filtering. The systems in Table 4 have 6.6% more words in their training than the corresponding systems in the D0 row in Table 5. The additional words contribute an absolute increase of 0.62 BLEU points (2.5% relative increase) in the ENR settings and a little below half of that in the RED settings.

5.4 Learning Curve Experiments

We also compare the value of different schemes across a learning curve where we consider smaller sets of our data : 2M, 1M and 0.5M words. We only show results for the reduced-then-enriched systems in Table 6. As expected, the increase in training data size causes an increase in BLEU scores. Both TB and S2 at the 2M level outperform D0 at the 4M level. The TB scheme is almost always the top performer. The S2 scheme goes from being ranked fourth in the smallest condition to being second in the largest. Further experiments considering the same learning curve with ENR training may be necessary to understand how different normalization settings interact with training size.

6 Conclusions and Future Work

We presented experiments studying a large number of variables for English-Arabic SMT systems that produce correctly tokenized and enriched Arabic text. The results shows that the lemma based alignment leads to a better output quality. Our best system uses the Penn Arabic Treebank (TB) tokenization scheme and reduced Arabic word forms followed by a language-model based joint detokenization and enrichment step. In the future we plan to investigate the use of system combination techniques and language modeling

	0.5M	1M	2M	4M
D0	19.73	22.26	24.04	24.66
D1	20.97	23.17	23.74	26.06
D2	21.33	**23.72**	24.21	26.06
TB	**21.74**	23.61	**25.22**	**26.73**
S2	20.62	23.04	24.80	26.11
D3	20.48	22.95	24.47	25.03

TABLE *6* – Comparing different tokenizations schemes over a learning curve using reduced-then-enriched systems

approaches that target Arabic's complex morphology such as factored LMs (Bilmes & Kirchhoff, 2003). We also plan to study the effect of using more data and explore other morphological tokenization schemes. Moreover, we plan to investigate the different types of MT errors in order to build detokenization systems that are robust to MT errors.

Acknowledgements

The work presented here was funded by a Google research award. We would like to thank Ioannis Tsochantaridis, Marine Carpuat, Alon Lavie, Hassan Al-Haj and Ibrahim Badr for helpful discussions.

Références

BADR I., ZBIB R. & GLASS J. (2008). Segmentation for english-to-arabic statistical machine translation. In *Proceedings of ACL-08 : HLT, Short Papers*, p. 153–156, Columbus, Ohio : Association for Computational Linguistics.

BADR I., ZBIB R. & GLASS J. (2009). Syntactic phrase reordering for English-to-Arabic statistical machine translation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, p. 86–93, Athens, Greece : Association for Computational Linguistics.

BILMES J. A. & KIRCHHOFF K. (2003). Factored language models and generalized parallel backoff. In *Proceedings of the Human Language Technology Conference/North American Chapter of Association for Computational Linguistics (HLT/NAACL-03)*, p. 4–6, Edmonton, Canada.

CALLISON-BURCH C., OSBORNE M. & KOEHN P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, p. 249–256, Trento, Italy.

EL KHOLY A. & HABASH N. (2010). Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proceedings of the seventh International Conference on Language Resources and Evaluation (LREC)*, Valletta, Malta.

ELMING J. & HABASH N. (2009). Syntactic reordering for English-Arabic phrase-based machine translation. In *Proceedings of the EACL 2009 Workshop on Computational Approaches to Semitic Languages*, p. 69–77, Athens, Greece : Association for Computational Linguistics.

GOLDWATER S. & MCCLOSKY D. (2005). Improving Statistical MT Through Morphological Analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 676–683, Vancouver, Canada.

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. & 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/Human Language Technologies Conference (HLT-NAACL06)*, p. 49–52, New York, NY.

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.

KOEHN P. (2004). Statistical significance tests formachine translation evaluation. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP'04)*, Barcelona, Spain.

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 *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, p. 177–180, Prague, Czech Republic.

LEE Y.-S. (2004). Morphological analysis for statistical machine translation. In *Proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04)*, p. 57–60, Boston, 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.

NIESSEN S. & NEY H. (2004). Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information. *Computational Linguistics*, **30**(2).

OCH F. J. & NEY H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(1), 19–52.

OFLAZER K. & DURGAR EL-KAHLOUT I. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, p. 25–32, Prague, Czech Republic : Association for Computational Linguistics.

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*, p. 311–318, Philadelphia, PA.

SARIKAYA R. & DENG Y. (2007). Joint morphological-lexical language modeling for machine translation. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Companion Volume, Short Papers*, p. 145–148, Rochester, New York : Association for Computational Linguistics.

STOLCKE A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, volume 2, p. 901–904, Denver, CO.

ZOLLMANN A., VENUGOPAL A. & VOGEL S. (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*, p. 201–204, New York City, USA : Association for Computational Linguistics.