# A Statistical Method for English to Arabic Machine Translation

Marwan Akeel
Department of computer Engineering
IIT-BHU, Varanasi, India

R. B. Mishra
Department of computer Engineering
IIT-BHU, Varanasi, India

## ABSTRACT
Translating from English into a morphologically richer language like Arabic is a challenge in statistical machine translation. Segmentation of Arabic text was introduced to bridge the inflection morphology gap. In this work, we investigate the impact of supporting Arabic morphologically segmented training corpus in a phrase-based statistical machine translation system with one to one dictionary and examine the effects on system performance. The results show that the dictionary improves the quality of the translation output especially when the corpus used is normalized and fully segmented excluding the determiner. The dictionary also decreases the out of vocabulary rate. The effect of the dictionary support with different baseline and factored models using data ranging from full word form to fully segmented forms are also demonstrated.

## General Terms
Machine translation.

## Keywords
Statistical machine translation, Factored phrase based, Natural language processing.

## 1. INTRODUCTION
Translations of a statistical machine translation (SMT) are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora. It has assumed that the word should be the basic token unit of translation which ignores any word internal morphological structure. Morphologically rich languages may produce a very large number of word forms for a given root form. Dealing with each form as a separate word leads us to large vocabulary growth, higher out-of-vocabulary rate and poor language model probability estimation for machine translation.

When it comes to Arabic, the importance of morphology is obvious. In a same size corpus, Arabic has more surface forms compared to any morphologically poor language like English. The morphological differences cause sparsity and ambiguity which affect the performance of the SMT. Several approaches have been proposed to minimize differences between Arabic and English. One of these approaches is word segmentation which becomes very important for many Natural Language Processing tasks that deal with Arabic language. The segmentation divides the Arabic words into its main components in order to make the English source align better with the segmented Arabic target. Striping a word from all the affixes and clitics makes it possible to replace it with other dictionary origin word which can be combined with the striped or different affixes and clitics to create different form of it which reduce the out-of-vocabulary rate.

Studies of Arabic morphological segmentation on SMT began with systems that translate from Arabic to English [12, 16, 25]. They found that reducing the sparsity caused by the rich morphology of Arabic improves the performance of Arabic to English SMT. [16] used trigram model to segment Arabic words and then delete or combine some of the segmented morphemes in order to enhance the alignment between the Arabic source and the English target. [12] proposed various segmentation schemes to segment the Arabic source. Both [12, 16] works showed that when the corpus size increases the benefit of segmentation decreases. The reason is that, with more training data the model becomes less sparse [4].

For Arabic as a target language, a growing number of studies has been published. [22] used joint morphological-lexical language models to re-rank the output of English-dialectal Arabic MT. [4] reported results on the value of morphological segmentation of Arabic during training. He also experimented on segmented factored data where he used the surface word, the stem and the POS tag concatenated to the segmented clitics and found that it performed better than the segmented phrase-based model but at a significantly higher cost in terms of time and required resources.

Some works were done on exploring Arabic segmentation schemes ranging from full word form to fully segmented forms and examining the effects on system performance. [1] found that a difference of 2.61 BLEU points between the best and worst segmentation schemes. [10] explored a space of tokenization schemes and normalization options.

Other works were done on unsegmentation/detokenization techniques. [4] described two different techniques for unsegmentation of Arabic in the output while [11] extended it and examined a set of six unsegmentation techniques over various segmentation schemes and compared two techniques for orthographic denormalization. [1] reported results on a wide set of techniques for combining the segmented Arabic output.

In this work we explore the benefit of supporting the Arabic training corpus, which ranges from full word form to fully segmented form, with one-to-one dictionary and examine the effects on system performance of both baseline phrase-based model and factored phrase based model. The factored used are POS and surface word for Arabic and only the surface word for English. The results show that supporting the training data with one to one dictionary benefits the quality of the translation specially when the data is fully segmented excluding the determiner (ال, Al) and benefits it most when the data is also normalized.

In the next section the Arabic orthography and morphological issues that affect the statistical machine translation are discussed. The approach used to tackle the issues is explained in the third section. Following that, the data used to conduct the experiments and its sizes is listed and then the experiments is described. Finally, the evaluation metrics and the results are discussed.

# 2. ORTHOGRAPHY AND MORPHOLOGICAL ISSUES

In this section, relevant aspects of Arabic word orthography and morphology which can affect the quality of SMT output is presented.

## 2.1 Arabic orthography

Arabic orthography influences the quality of the output translation of the SMT systems when the Arabic training data is not correctly orthographically written. Some of the issues and common mistakes committed by a regular Arabic writer are listed below:

- The incomplete or missing of diacritics which are used in the Arabic writing system to represent the short vowels.

Foundation: الْمُؤَسَّسَةُ - الْمُؤَسَّسة - المؤسسة

- The drop of hamza ( ء ). In particular, variants of HamzatedAlif ( أ, إ ) where it is commonly written without their Hamza ( ا ). The three forms of Alifare often used interchangeably.

Smile: ابتسامة - إبتسامة - أبتسامة

- Two dots inserted on Aleph Maqsura ( ى ), and two dots removed from Yaa ( ي )in word final position.

Until: حتّي - حتى

My pen: قلمى - قلمي

- Two dots inserted on final Haa ( ه ), and two dots removed from TaaMarbouta ( ة )

Smile: ابتسامه - ابتسامة

Waters: مياه - مياة

The above mentioned orthography issues produce:

- Multiple forms of the same word which increase the sparsity.
- Same form corresponding to multiple words which increase the ambiguity.

## 2.2 Morphology

Arabic is morphologically rich with highly complex word formation of roots and patterns producing a large number of rich word forms. An Arabic word may be constructed out of a stem plus affixes and clitics. Inflectional affixes are used on verbs to encode person, number, gender, tense, and mood information, and on nouns to encode gender, number, and case information. A clitic is a linguistic unit attached to a stem written and pronounced like an affix but it is grammatically independent. Some of the proclitics and enclitics used in Arabic are shown in table1, where enclitics are marked by "+"at the beginning and proclitics are marked by "#" at the end. Except for the definite article, all the clitics listed in table 1 are extracted from [17]**.** Arabic transliteration are provided in Buckwalter transliteration scheme [5].

The heavy existence of clitics in Arabic increases the lexicon size. For a similar broad linguistic content, Arabic needs a lexicon of a size equivalent to 1.76 times of an English lexicon [3].

These Arabic various attachable clitics cause sparsity, alignment and matching issues with languages that have very little morphology like English. According to [10], while the number of Arabic words in a parallel corpus is 20% less than English words, the number of unique Arabic words is over 50% more than the number of unique English words.

**Table 1. Clitics in Arabic**

| Clitics | Transliteration | category |
|---|---|---|
| ال# | Al# | Definite Article |
| و# | w# | Conjunction, coordinating |
| ف# | f# | Conjunction, subordinating |
| ل# | l# | Preposition |
| ب# | b# | Preposition |
| ك# | k# | Preposition |
| س# | s# | Future verbal particle |
| +ي | +y | POSS_PRON_1S/ PRON_1S |
| +ا | +A | PRON_1P |
| +ني | +ny | IVSUFF_DO/PRON_1S/PVSUFF_DO |
| +ك | +k | POSS_PRON_2MS/ PRON_2MS |
| +كما | +kmA | POSS_PRON_2D/ PRON_2D |
| +كم | +km | POSS_PRON_2MP/ PRON_2MP |
| +كن | +kn | POSS_PRON_2FP/ PRON_2FP |
| +ه | +h | POSS_PRON_3MS/ PRON_3MS |
| +ها | +ha | POSS_PRON_3FS/ PRON_3FS |
| +هما | +hmA | POSS_PRON_3D/ PRON_3D |
| +هن | +hn | POSS_PRON_3FP/ PRON_3FP |
| +هم | +hm | POSS_PRON_3MP/ PRON_3MP |
| +نا | +nA | POSS_PRON_1P/ PRON_1P |

# 3. APPROACH

To tackle the issues of orthography and morphology mentioned in the previous section the Arabic data has to be preprocessed before the training and the resulted output should be post processed later on.

## 3.1 Preprocessing

Before working on the data, the Arabic and English Data were aligned, cleaned and lowercased. The sentences that are more than 100 tokens and the empty lines were removed. The data was also tokenized by separating the words from punctuations and numbers. The data was then segmented and factorized. For the issues of orthography mentioned above, only the Hamzated Alif was normalized while the other cases was left as it is ( Ta Marbouta and Haa, Aleph Maqsura and Yaa) because they affect the output quality of segmentation and unsegmentaion processes we followed in this work. The Hamzated Alif is normalized by changing its various forms ( أ , إ ) to bare Alif ( ا ). All the diacritics were removed wherever they occur.

**Table 2. Segmentation schemes example.**

| Scheme | Segmentation and tagging example |
|---|---|
| Full form (FF) | • وبسيارته سيأخذونها للنزهه<br>• wbsyArth sy>x*wnhA llnzhh |
| | • وبسيارته\|NN سيأخذونها\|VBP للنزهه\|NN<br>• wbsyArth\|NN sy>x*wnhA\|VBN llnzhh\|NN |
| Fully Tokenized (FT) | • و# ب# سيارة +ه س# يأخذون +ها ل# ال# نزهه<br>• w# b# syArtp +h s# y>x*wn +hA l# Al# nzhh |
| | • و#\|CC ب#\|IN سيارة\|BD_FS3 +ه\|PRP_MS3 س#\|FP يأخذون\|VBP_MP3 +ها\|PRP_FS3 ل#\|IN ال#\|DET نزهه\|NN<br>• w#\|CC b\|IN syArp\|VBD_FS3 +h\|PRP_MS3 s#\|FP y>x*wn\|VBP_MP3 +hA\|PRP_FS3 l#\|IN Al#\|DET nzhh\|NN |
| CPF | • و# ب# سيارته س# يأخذونها ل# النزهه<br>• w# bsyArth s# y>x*wnhA l# Al# nzhh |
| | • و#\|CC ب#\|IN سيارته\|VBD_FS3_PRP_MS3 س#\|FP يأخذونها\|VBP_MP3_PRP_FS3 ل#\|IN النزهه\|DET_NN<br>• w#\|CC b\|IN syArth\|VBD_FS3_ PRP_MS3 s#\|FP y>x*wnhA\|VBP_MP3_PRP_FS3 l#\|IN Al#\|DET nzhh\|NN |
| Prefix | • و# ب# سيارته س# يأخذونها ل#ال# نزهه<br>• w# bsyArth s# y>x*wnhA l#Al# nzhh |
| | • و#\|CC ب#\|IN سيارته\|VBD_FS3_PRP_MS3 س#\|FP يأخذونها\|BP_MP3_PRP_FS3 ل#ال#\|IN_DET نزهه\|NN<br>• w#\|CC b\|IN syArth\|VBD_FS3_PRP_MS3 s#\|FP y>x*wnHA\|VBP_MP3_PRP_FS3 l#Al#\|IN_DET nzhh\|NN |
| CPFSuff | • و# ب# سيارة +ه س# يأخذون +ها ل# النزهه<br>• w# b# syArp +h s# y>x*wn +hA l# Alnzhh |
| | • و#\|CC ب#\|IN سيارة\|BD_FS3 +ه\|PRP_MS3 س#\|FP يأخذون\|VBP_MP3 +ها\|PRP_FS3 ل#\|IN النزهه\|DET_NN<br>• w#\|CC b#\|IN syArp\|VBD_FS3 +h\|PRP_MS3 s#\|FP y>x*wn\|VBP_MP3 +hA\|PRP_FS3 l#\|IN Alnzhh\|DET_NN |
| Suff | • وبسيارت +ه سيأخذون +ها للنزهه<br>• wbsyArt +h sy>x*wn +hAllnzhh |
| | • وبسيارت\|CC_VBD_FS3 +ه\|PRP_MS3 سيأخذون\|FP_VBP_MP3 +ها\|PRP_FS3 للنزهه\|IN_DET_NN<br>• wbsyArt\|CC_VBD_FS3 +h\|PRP_MS3 sy>x*wn\|FP_VBP_MP3 +hA\|PRP_FS3 llnzhh\|IN_DET_NN |

For the morphological issues, the clitics was extracted out of the stem by segmenting the training, decoding, tuning and testing data. Five types of segmentation schemes have been applied which differ on the clitics participate in segmentation.

To perform pre-translation morphological segmentation of the Arabic source, AMIRA 2.0 toolkit was used [7]. The AMIRA toolkit includes a clitic tokenizer, part of speech tagger (POS) and base phrase chunker - shallow syntactic parser. The AMIRA system does not handle inflectional morphology. The tokenization system has an F score measure of 99.2%. Both POS taggers with their different POS tag sets perform at over 96% accuracy.

The factors on the Arabic side are the POS tags and the surface word. On the English side, only the surface word is used. The full form Arabic source is tagged with POS using Stanford POS Tagger [24], while AMIRA toolkit is used to tag the segmented Arabic data with POS.

For all the different factored and baseline phrase-based models experiments, the same Arabic source was used but with different forms, five segmentation schemes in addition to the origin (unsegmented) form. The five segmentation schemes are:

1. Fully tokenized (FT): conjunctions, prepositions, determiners, suffixes and future markers are all individually separated.
2. CPF: conjunctions, prepositions and future markers are all individually separated.
3. Prefix: All prefixes are separated as one token.
4. CPFSuff: conjunctions, prepositions, future markers and suffixes are all individually separated.
5. Suff : Only suffixes are separated.

Table 2 demonstrate an example of these segmentation schemes.

## 3.2 Post processing

The Arabic output produced by all SMT models that uses segmented Arabic corpus needs to be recombined in order to produce the final Arabic text. This step is called unsegmentation/detokenization.

A technique uses manually defined morphological adjustments rules were applied to combine the Arabic segments. Some of these rules are shown on table 3.

**Table 3. Some of the manually defined morphological adjustments rules.**

| Rule | Example (right to left) |
|---|---|
| " ة " + pron/poss → " ت " + pron/poss | Their city "مدينة" + "+هم" ← "مدينتهم" |
| " ى " + pron/poss → " ا " + pron/poss | He built it "بنى" + "+ها" ← "بناها" |
| " ال " + " ل# " → "لل" | for the security "ل#" + "الأمن" ← "للأمن" |
| " ال " + " ال# " → " الل " | The refugees "ال#" + "الاجئين" ← " اللاجئين " |

# 4. DATA USED

The English-Arabic parallel training data was collected from different sources in the internet. These data are from culture, economy, politics, religion and sports. Some data were collected manually, sentence by sentence, to support the training data with poetry, literature and technological vocabulary. The size of the training data is illustrated in table

4. The data was arranged, cleaned and parallelized to fit into the training tool.

The one-to-one dictionary used to support the training data is called Ekseer Dictionary[9]. It has around 95 thousands one-to-one words. The dictionary was cleaned and modified to keep only one translation for each entry.

**Table 4. Training corpus size.**

| Corpus | No. of sentences | No. of Arabic words | No. of English words |
|---|---|---|---|
| Holy Quran[13] | 6236 (3 references) | 77.8K | 155K 161K 167K |
| Meedan-Memory [18] | 18732 | 413K | 433K |
| UN corpus[2] | 74423 | 2.68M | 3M |
| Manually Collected (from different web pages) | 5698 | 82.5K | 108K |
| **TOTAL** | 105K | 3.254M | 4.06M |

The Language model Arabic corpus contains of the Arabic side of the training data plus data from news papers and articles (see table 5). Tuning was done using 480 parallel sentences and the model was tested on 303 sentences from different domains.

**Table 5. Language model corpus size.**

| Corpus | No. of Words |
|---|---|
| Khaleej-2004 (news paper) | 2.5M |
| Watan-2004 (news paper) | 10M |
| Latifa Al-Sulaiti collection [15] | 700K |
| Training data (Arabic side) | 3.254M |
| TOTAL | 16.5M |

# 5. EXPERIMENTS

All experiments were conducted using GIZA++ [19] to align the English source to the segmented/unsegmented Arabic. The decoding is done using MOSES toolkit [14] with grow-diag-final-and heuristic to symmetrize the alignement, msd-bidirectional-f reordering model and Good Turing as score option. The value given to the maximum length of phrases entered into phrase table varies according to the increase of word count of each scheme after segmentation. Table6 demonstrates the increasing rate of word count of each scheme and the value given to the max-phrase-length where the default value 7 is given to the unsegmented corpus. Tuning is done using Och's algorithm [19]. 5-grams were used for all surface words language models and 7-grams for the POS language models. All language models were implemented using the SRILM toolkit [23].

Forty eight experiments were conducted which differed from one another in the type of data used. The data used is a combination of baseline (B) or factored (F), normalized (N) or raw (R), with Dictionary support (D) or without the dictionary support, and segmented or full form (FF).

**Table 6. max-phrase-length value for different segmentation schemes.**

| Scheme | Rate of increasing of word count | Value of max-phrase-length |
|---|---|---|
| FF | - | 7 |
| FS | 44% | 10 |
| CPF | 15% | 8 |
| Prefix | 37% | 10 |
| CPFSuff | 21% | 9 |
| suff | 6% | 7 |

# 6. EVALUATION AND RESULTS

The outputs were evaluated using BLEU [20], NIST [8] and METEOR [6] automatic metrics. BLEU reports high correlation with human judgment of quality and is one of the most popular metric in the field. It calculates scores for individual segments, generally sentences, and then averages these scores over the whole corpus for a final score. NIST is based on BLEU metric but with some alteration. Where BLEU simply calculates n-gram precision adding equal weight to each one, NIST also calculates how informative a particular n-gram is by giving more weight for correct rarer n-gram found on the translation and lower weight for more likely occurring n-gram. METEOR is designed to address some of the deficiencies inherent in the BLEU metric. It includes synonymy matching, where instead of matching only on the exact word form, the metric also matches on synonym. It also includes a stemmer, which lemmatises words and matches on the lemmatised forms.

The scores of the forty eight experiments are shown in table 7. The increase and decrease of the score rate are measured based on the baseline model scores which uses unprocessed full form raw corpus (FF + B + R, no. 1) which is listed at the top of the table. The rate value is calculated by subtracting the base model metrics scores from each model metrics scores individually and then multiplying the result by 100 in case of BLEU and METEOR, and by 10 in case of NIST because the BLEU and METEOR scores are between 0 and 1 while NIST score is between 0 and 10.

In general, the baseline segmented models perform better than the factored segmented models. Looking at table 7 and figure 1, The best METEOR and NIST metrics translation scores are obtained from the FS, baseline, normalized, with dictionary support model (FS+B+N+D, no. 12), which have improved scores by 4.89% and 1.90% respectively. The best BLEU score isobserved at the CPF, baseline, normalized, with dictionary support model (CPF+B+N+D, no. 36) where the improvement in BLEU score is 3.19%.

The worst scores are generally found at Prefix models. The lowest scores are obtained from the factored, suffix separated, and raw data model (Suff+F+R, no. 45) where all metrics scores are negative, and the METEOR and NIST are at their lowest value showing that the translation lacks the support of the vocabulary and synonymy. It is important to keep in mind that the accuracy of the tokenizer is not totally perfect especially for Prefix and CPF segmentation schemes where many Prefix and CPF cases were not segmented causing sparsity and affecting the scoring. The worst BLEU score

**Table 7. Experiments output scores.**

| Serial no. | Phrase-based Data Type | NIST Score<br>1%=0.1NIST score | BLEU Score<br>1%=0.01BLEU score | METEOR Score<br>1%=0.01METEOR score |
|---|---|---|---|---|
| 1. | FF + B + R | 3.5094 | 0.1037 | 0.0864 |
| 2. | FF + B + R + D | +2.21% | +0.24% | +0.68% |
| 3. | FF + B + N | +1.18% | +0.79% | +0.80% |
| 4. | FF + B + N +D | +3.49% | +0.57% | +1.44% |
| 5. | FF + F + R | -1.17% | -0.79% | -0.13% |
| 6. | FF + F + R + D | +0.87% | +0.29% | +0.23% |
| 7. | FF + F + N | +0.99% | +0.69% | +0.86% |
| 8. | FF + F + N + D | +2.74% | +0.93% | +1.27% |
| 9. | FS + B + R | +0.86% | -0.16% | +0.44% |
| 10. | FS + B + R + D | +2.87% | +0.67% | +1.02% |
| 11. | FS + B + N | +2.29% | +1.22% | +1.39% |
| 12. | FS + B + N + D | +4.89% | +2.39% | %+1.90 |
| 13. | FS + F + R | -0.20% | -1.23% | +0.18% |
| 14. | FS+ F + R + D | +1.06% | -0.95% | +0.40% |
| 15. | FS + F + N | +0.76% | -0.77% | +0.81% |
| 16. | FS + F + N + D | +2.34% | +0.53% | +1.23% |
| 17. | CPFSuff + B + R | +0.26% | +0.22% | +0.28% |
| 18. | CPFSuff + B + R + D | +1.67% | +0.57% | +0.34% |
| 19. | CPFSuff + B + N | +1.05% | -0.33% | +0.83% |
| 20. | CPFSuff + B + N + D | +4.78% | +3.08% | +1.71% |
| 21. | CPFSuff + F + R | -0.59% | -1.17% | +0.00% |
| 22. | CPFSuff + F + R + D | +1.91% | +1.49% | +0.56% |
| 23. | CPFSuff + F + N | +1.49% | +0.39% | +0.98% |
| 24. | CPFSuff + F + N + D | +3.35% | +2.21% | +1.27% |
| 25. | Prefix + B + R | -0.08% | -0.02% | +0.06% |
| 26. | Prefix + B + R + D | +0.89% | -0.07% | +0.48% |
| 27. | Prefix + B + N | +1.30% | +0.30% | +0.86% |
| 28. | Prefix + B + N + D | +2.96% | +0.40% | +0.84% |
| 29. | Prefix + F + R | -0.88% | -2.06% | -0.09% |
| 30. | Prefix + F + R + D | -0.72% | -1.79% | +0.53% |
| 31. | Prefix + F + N | -0.73% | -1.54% | +0.05% |
| 32. | Prefix + F + N + D | +1.38% | -0.01% | +1.14% |
| 33. | CPF + B + R | -0.60% | -0.96% | -0.05% |
| 34. | CPF + B + R + D | +1.31% | +0.43% | -0.16% |
| 35. | CPF + B + N | +3.07% | +1.4% | +1.17% |
| 36. | CPF + B + N + D | +3.72% | +3.19% | +1.04% |
| 37. | CPF + F + R | +1.60% | +1.74% | +0.31% |
| 38. | CPF + F + R + D | +2.48% | +1.95% | +0.23% |
| 39. | CPF + F + N | +2.91% | +1.96% | +1.13% |
| 40. | CPF + F + N + D | +3.55% | +0.88% | +0.80% |
| 41. | Suff + B + R | +0.12% | -0.45% | +0.18% |
| 42. | Suff + B + R + D | +1.88% | -0.36% | +0.29% |
| 43. | Suff + B + N | +2.38% | +0.63% | +1.04% |
| 44. | Suff + B + N + D | +2.76% | +0.64% | +0.96% |
| 45. | Suff + F + R | -1.90% | -1.14% | -0.16% |
| 46. | Suff + F + R + D | +0.11% | -0.39% | +0.26% |
| 47. | Suff + F + N | +2.30% | +1.04% | +1.36% |
| 48. | Suff + F + N + D | +3.45% | +0.33% | +1.43% |

obtained when apply the Prefix scheme on factored and raw data (Prefix+F+R, no. 29).

Looking at the experiments which are supported with dictionary, it is found that it has the best scores in general compared to similar experiments with no dictionary support (every experiment with even number out performs the preceding experiment with odd number). The benefit of using dictionary is well-recognized at the baseline segmented models. The benefit of using one-to-one dictionary support is gained when the data is normalized and segmented. The best segmentation scheme that benefits from using one-to-one dictionary is the CPFSuff, where the prefixes and suffixes are separated except the determiners. The CPFSuff+B+N+D model scores in NIST, BLEU and METEOR increased by 3.73%, 2.75%, and 0.88% respectively compared to CPFSuff+B+N model. Table 7 also shows that Suff scheme has the lowest benefit of the dictionary support. CPFSuff and

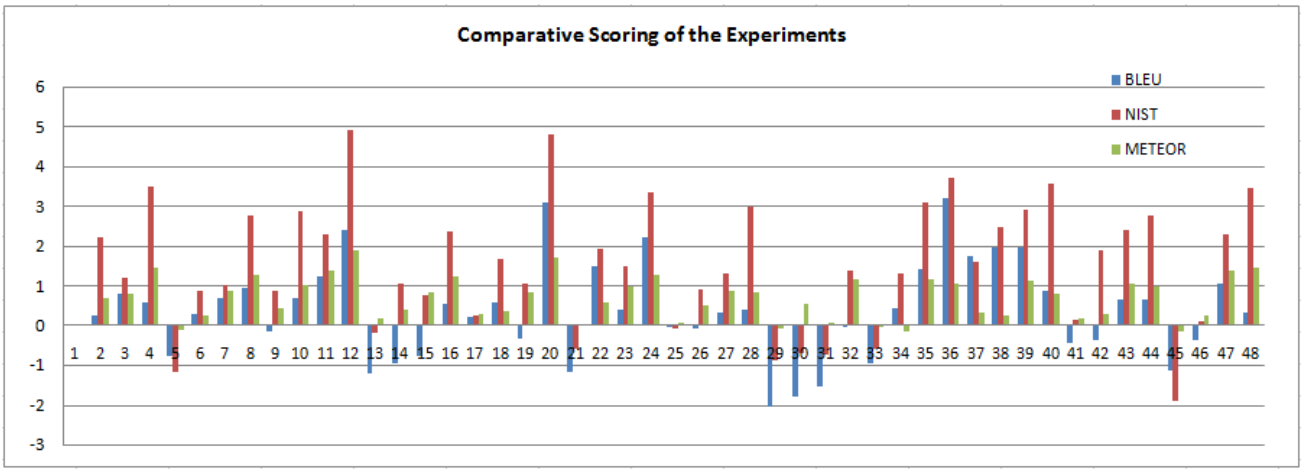


**Fig 1: Comparative scoring of the experiments in table 7.**

Prefix models would have scored better if the tokenizer had been more accurate.

## 7. CONCLUSION

Segmenting the Arabic text has improved the SMT output for both the cases where Arabic is the target or the source. The advantage of the ability of segmenting the Arabic text was utilized and support the training corpus with one-to-one dictionary. It was found that it has benefited the quality of the translation output. Different segmentation schemes were experimented with and without dictionary support. Considering the NIST, Bleu and Meteor scores, it is found that the base line, normalized, fully segmented with dictionary support form is the most suitable scheme to benefit from the dictionary support and segmentation where NIST and Meteor metrics are in their highest. Blue metric scored its highest at the scheme that segments only the conjunction, preposition and future marks with dictionary support and normalized text. Due to the higher cost, in term of time and required resources, and the big number of conducted experiments (forty eight experiments), 105K sentence pair of training data and two factors with the Arabic corpus was used. [4, 16, 21] works show that the improvements obtained from segmentation decrease as the corpus size increase , which is due to the fact that the model becomes less sparse with more training data, so the data used is sufficient.

## 8. REFERENCES

[1] Al-Haj, Hassan and Lavie, Alon. The impact of Arabic morphological segmentation on broad-coverage English-to-Arabic statistical machine translation. *Machine translation*, vol. 26, pp. 3-24, 2012.

[2] Alexandre Rafalovitch, Robert Dale. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. in *Proceedings of the MT Summit XII*, pp. pages 292-299, Ottawa, Canada. CiteULike record for the paper (UN REFERENCE), 2009.

[3] Alotaiby, Fahad, Alkharashi, Ibrahim, and Foda, Salah. Processing large Arabic text corpora: Preliminary analysis and results. in *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, pp. 78-82, 2009.

[4] Badr, Ibrahim , Zbib, Rabih, and Glass, James. Segmentation for English-to-Arabic statistical machine translation. presented at the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers, Columbus, Ohio, 2008.

[5] Buckwalter, Tim. Buckwalter Arabic Morphologcal Analyzer. in *Linguistic Data Consortium. (LDC2002L49)*, 2002.

[6] Denkowski, Michael and Lavie, Alon. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. in *Proceedings of the EMNLP 2011 Workshop on Statistical Machine Translation*, pp. 85-91, 2011.

[7] Diab, Mona. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. in *2nd International Conference on Arabic Language Resources and Tools*, 2009.

[8] Doddington, George. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. in *Proceedings of the second international conference on Human Language Technology Research*, pp. 138-145, 2002.

[9] Ekseer Dictionary. http://at.alixsys.com/codesprint2009/sandbox/taha/ مشروع القاموس/Ekseer Dictionary.mdb, last accessed on (2012, May).

[10] El Kholy, Ahmed and Habash, Nizar. Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. *TALN 2010, Montréal*, 2010.

[11] El Kholy, Ahmed and Habash, Nizar. Techniques for Arabic morphological detokenization and orthographic denormalization. in *Workshop on Language Resources and Human Language Technology for Semitic Languages in the Language Resources and Evaluation Conference (LREC), Valletta, Malta*, 2010.

[12] Habash, Nizar and Sadat, Fatiha. Arabic preprocessing schemes for statistical machine translation. in *In Proc. of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pp. 49–52, New York City, NY, 2006.

[13] Knight, Kevin, Al-Onaizan, Yaser, Purdy, David, Curin, Jan, Jahr, Michael, Lafferty, John, Melamed, Dan, Smith, Noah, Och, Franz Josef, and Yarowsky, David. EGYPT: a statistical machine translation toolkit. http://old-site.clsp.jhu.edu/ws99/projects/mt/, last accessed on (1999, Nov 2012).

[14] Koehn, Philipp, Hoang, Hieu, Birch, Alexandra, Callison-Burch, Chris, Federico, Marcello, Bertoldi, Nicola, Cowan, Brooke, Shen, Wade, Moran, Christine,

and Zens, Richard. Moses: Open source toolkit for statistical machine translation. in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 177-180, 2007.

[15] Latifa, Al-Sulaiti http://www.comp.leeds.ac.uk/latifa/research.htm, last accessed on (2012, 12 Nov.).

[16] Lee, Young-Suk. Morphological analysis for statistical machine translation. in *Proceedings of the 5th Meeting of the North American Chapter of the Association for Computational Linguitics/Human Language Technologies Conference (HLT NAACL04)*, pp. 57–60, Boston, MA, 2004.

[17] Maamouri, M., Bies, A., Kulick, S., Gaddeche, F., and Mekk, W. Arabic Treebank: Part 3(a) v. 2.6. presented at the Linguistic Data Consortium. , Philadelphia, USA, Catalog ID: LDC2007E65., 2007.

[18] Meedan. Meedan's Open Source Arabic/English Translation Memory

http://github.com/anastaw/Meedan-Memory, last accessed on (2012, Sep.).

[19] Och, Franz Josef and Ney, Hermann. A systematic comparison of various statistical alignment models. *Computational linguistics*, vol. 29, pp. 19-51, 2003.

[20] Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing. BLEU: a method for automatic evaluation of machine translation. in *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311-318, 2002.

[21] Sadat, Fatiha and Habash, Nizar. 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 ACL (Coling ACL'06)*, pp. 1-8, Sydney, Australia, 2006.

[22] Sarikaya, Ruhi and Deng, Yonggang. 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*, pp. 145-148, 2007.

[23] Stolcke, Andreas. SRILM-an extensible language modeling toolkit. in *Proceedings of the international conference on spoken language processing*, pp. 901-904, 2002.

[24] Toutanova, Kristina, Klein, Dan, Manning, Christopher D, and Singer, Yoram. Feature-rich part-of-speech tagging with a cyclic dependency network. in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pp. 173-180, 2003.

[25] Zollmann, Andreas, Venugopal, Ashish, and Vogel, Stephan. Bridging the inflection morphology gap for Arabic statistical machine translation. presented at the Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, New York, New York, 2006.