# Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language

**Nizar Habash**
Center for Computational Learning Systems
Columbia University
New York, NY 10115, USA
habash@ccls.columbia.edu

**Jun Hu**
Computer Science Department
Columbia University
New York, NY 10115, USA
jh2740@columbia.edu

## Abstract

We present a comparison of two approaches for Arabic-Chinese machine translation using English as a pivot language: sentence pivoting and phrase-table pivoting. Our results show that using English as a pivot in either approach outperforms direct translation from Arabic to Chinese. Our best result is the phrase-pivot system which scores higher than direct translation by 1.1 BLEU points. An error analysis of our best system shows that we successfully handle many complex Arabic-Chinese syntactic variations.

## 1 Introduction

Arabic and Chinese are two languages with a very large global presence; however, there has not been, to our knowledge, any work on MT for this pair. Given the cost involved in creating parallel corpora for Arabic and Chinese and given that there are lots of available resources (in particular parallel corpora) for Arabic and English and for Chinese and English, we are interested in exploring the role English might serve as a pivot (or bridge) language. In this paper we explore different ways of pivoting through English to translate Arabic to Chinese. Our work is similar to previous research on pivot languages except in that our three languages (source, pivot and target) are very different and from completely unrelated families. We focus our experiments on a trilingual parallel corpus to keep all conditions experimentally clean. Our results show that using English as a pivot language for translating Arabic to Chinese actually outperforms direct translation. We believe this may be a result of English being a sort of middle ground between Arabic and Chinese in terms of different linguistic features and, in particular, word order.

Section 2 describes previous work. Section 3 discusses relevant linguistic issues of Arabic, Chinese and English. Section 4 describes our system and different pivoting techniques. And Section 5 presents our experimental results.

## 2 Previous Work

There has been a lot of work on translation from Chinese to English (Wang et al., 2007; Crego and Mariño, 2007; Carpuat and Wu, 2007; among others) and from Arabic to English (Sadat and Habash, 2006, Al-Onaizan and Papineni, 2006; among others). There is also a fair amount of work on translation into Chinese from Japanese, Korean and English (Isahara et al., 2007; Kim et al., 2002; Ye et al., 2007; among others). In 2008, the National Institute of Standards and Technology (NIST) MT Evaluation competition introduced English-Chinese as a new evaluation track.[1]

Much work has been done on exploiting multilingual corpora for MT or related tasks such as lexical induction or word alignment. Schafer and Yarowsky (2002) induced translation lexicons for languages without common parallel corpora using a bridge language that is related to the target languages. Simard (1999) described a sentence aligner that makes simultaneous decisions in a trilingual parallel text. Kumar et al. (2007) improved Arabic-English MT by using available parallel data in other languages. Callison-Burch et al (2006) exploited the existence of multiple parallel corpora to learn paraphrases for Phrase-based MT. Filali and Bilmes (2005) improved word alignment by leveraging multilingual parallel translations.

Most related to our work on pivoting are the following: Utiyama and Isahara (2007) studied

[1] http://www.nist.gov/speech/tests/mt/2008/doc/

*Proceedings of the Fourth Workshop on Statistical Machine Translation* , pages 173–181,
Athens, Greece, 30 March – 31 March 2009. 

sentence and phrase pivoting strategies using three European languages (Spanish, French and German). Their results showed that pivoting does not work as well as direct translation. Wu and Wang (2007) focused on phrase pivoting. They proposed an interpolated scheme that employs two phrase tables: one extracted from a small amount of direct parallel data; and the other extracted from large amounts of indirect data with a third pivoting language. They compared results for different European language as well as Chinese-Japanese translation using English as a pivoting language. Their results show that simple pivoting does not improve over direct MT; however, extending the direct MT system with phrases learned through pivoting helps. Babych et al. (2007) compared two methods for translating into English from Ukrainian: direct Ukrainian-English MT versus translation via a cognate language, Russian. Their comparison showed that it is possible to achieve better translation quality via pivoting.

In this paper we use a standard phrase-based MT approach (Koehn, 2004) that is in the same spirit of most statistical MT nowadays. We believe that we are the first to explore the Arabic-Chinese language pair in MT. We differ from previous pivoting research in showing that pivoting can outperform direct translation even when the source, target and pivot languages are all linguistically unrelated.

## 3 Linguistic Issues

In this section we discuss different linguistic phenomena in which Arabic, English and Chinese are divergent. We consider orthography, morphology and syntax. We also present a new metric for quantifying linguistic differences.

### 3.1 Orthography

Arabic is written from right-to-left using an alphabet of 36 letters and eight *optional* diacritical marks. Arabic is written in a cursive mostly word-internal connected form, but words are separated by white spaces. The absence of Arabic diacritics adds a lot of ambiguity.

Chinese uses a complex orthography that includes around 10,000 characters in common use. Characters convey semantic rather than phonological information. Chinese is written from left-to-right or top-down. Chinese words can be made out of one, two or more characters. However, words are written without separating spaces. Word segmentation is a major challenge for processing Chinese (Wu, 1998).

English uses the Roman alphabet and its words are written with separating white spaces. English orthography is much closer to Arabic than it is to Chinese.

### 3.2 Morphology

Arabic is a morphologically rich language with a large set of morphological features such person, number, gender, voice, aspect, mood, case, and state. Arabic features are realized using both concatenative (affixes and stems) and templatic (root and patterns) morphology with a variety of morphological, phonological and orthographic adjustments. In addition, Arabic has a set of very common clitics that are written attached to the word, e.g., the conjunction +و *w*+ 'and', the preposition +ب *b*+[2] 'with/in', the definite article +ال *Al*+ 'the' and a range of pronominal clitics that can attach to nouns (as possessives) or verbs and prepositions (as objects).

In stark contrast to Arabic, Chinese is an isolating language with no morphology to talk of. However, what Chinese lacks in morphology it replaces with a complex system of nominal quantifiers and verbal aspects. For example, in Figure 1 (at the end of this paper), Chinese marks the definiteness and humanness of the word 学生 *Xue Sheng* 'student' using the two characters 这位 *Zhe Wei* 'this person', while the indefiniteness and book-*ness* of the word 书 *Shu* 'book' are indicated through the characters 一本 *Yi Ben* 'one book-type'.

English has a simple limited morphology primarily indicating number and tense. English stands in the middle between Arabic and Chinese in terms of morphological complexity.

### 3.3 Syntax

Arabic is morpho-syntactically complex with many differences from Chinese and English. We describe here three prominent syntactic issues in which Arabic, Chinese and English vary widely: subject-verb order, verb-

[2] Arabic transliteration is in the Habash-Soudi-Buckwalter scheme (Habash et al. 2007).

prepositional phrase order and nominal modification.

First, Arabic verb subjects may be: (a.) pro-dropped (verb conjugated), (b.) pre-verbal, or (c.) post-verbal. The morphology of the Arabic verb varies in the three cases. By contrast English and Chinese are both generally subject-verb languages. When translating from Arabic, the challenge is to determine whether there is an explicit subject and, if there is, whether it is pre- or post-verbal. Since Arabic objects also follow the verb, a sequence of *Verb NounPhrase* may be a *verb subject* or a *pro-drop-verb object*. In the example in Figure 1, the subject (*student*) appears after the sentence initial verb in Arabic, but at the beginning of the sentence in Chinese and English.

Secondly, as for the word order of prepositional phrases (PP), Arabic and English are similar in that PPs generally appear at the end of the sentence (after all the verbal arguments) and to a lesser extent at its beginning. In Chinese, however, some PPs (in particular locatives and temporals) must appear between subject and verb. Other PPs may appear at end of sentence. In the example in Figure 1, the location of the *reading*, 'in the classroom' appears at the end of the Arabic and English sentences; however, it is between subject and verb in Chinese.

Finally, we distinguish three types of nominal modification: adjectival (as in 'red book'), possessive (as in 'John's book') and relative (as in 'the book [which] John gave me'). All of these modification types are handled in a similar manner in Chinese: using the particle 的 *De* to connect modifier with modified. Modifiers always precede the modified. For example, in Figure 1, 'a book about China' appears as 关于 中国 的书 *Guan Yu Zhong Guo De Shu 'about China De book'*. Similarly, 'the student's book' would be translated as 学生 的 书 *Xue Sheng De Shu 'student DE book'*. Like Chinese, English adjectival modifiers precede what they modify. However, relative modifiers follow. Possessive modifiers in English can appear before or after: 'the student's book' or 'the book of the student'. Unlike English and Chinese, Arabic adjectival modifiers typically follow their nouns (with a small exception of some superlative adjectives). However, similar to English but not Chinese, Arabic relative modifiers follow what they modify. As for possessive modifiers, Arabic has a special construction called *Idafa*, in which modifiers immediately follow what they modify without connecting particles. For example, 'the student's book' can only be translated in Arabic as كتاب الطالب *ktAb AlTAlb 'book the-student'*.[3]

These different phenomena are summarized in Table 1. It is interesting to point out that English phenomena are a middle ground for Arabic and Chinese: in some cases English is closer to Arabic and in others to Chinese.

| | **Arabic** | **English** | **Chinese** |
|---|---|---|---|
| **Orthography** | reduced alphabet | alphabet | Characters |
| **Morphology** | Rich | Poor | Very Poor |
| **Subject-Verb** | V Subj<br>Subj V<br>$V_{subj}$ | Subj V | Subj … V |
| **Verb-PP** | V…PP | V…PP | PP V<br>V PP |
| **Adjectival Modifier** | N Adj | Adj N | Adj DE N |
| **Possessive Modifier** | N Poss | N of Poss<br>Poss 's N | Poss DE N |
| **Relative Modifier** | N Rel | N Rel | Rel DE N |

Table 1: Comparing different linguistic phenomena in Arabic, English and Chinese

### 3.4 Quantifying Linguistic Differences

The previous section described specific types of linguistic phenomena without distinguishing them in terms of frequency or effect distance. For example, Arabic nominals (nouns, adjectives and adverbs) are seven times as frequent as verbs; and nominal modification phenomena are more likely local than long distance compared to verb-subject order. A proper quantification of these different phenomena requires trilingual parallel treebanks, which are not available. As such, we propose a simple metric to quantify linguistic differences by measuring the translation complexity of different language pairs. The metric is Average Relative Alignment Length (ARAL):

$$ARAL = \frac{1}{|L|} \sum_{l_{ab} \in L} \left| \frac{p_a}{S_a} - \frac{p_b}{S_b} \right|$$

[3] Arabic dialects allow an additional construction. We focus here on Modern Standard Arabic.

We define $L$ as the set of all alignment links linking words in a parallel corpus of languages A and B. For each alignment link, $l_{ab,}$ linking words $a$ and $b$, we define $p_a$ and $p_b$ as the position of words $a$ and $b$ in their respective sentences. We also define $S_a$ and $S_b$ as the lengths of the sentences in which $a$ and $b$ appear, respectively. ARAL is the mean of the absolute difference in relative word position ($p_i/S_i$) of the words of every alignment link. The larger ARAL is, the more reordering and insertions/deletions we expect, and the more complexity and difference. ARAL is a harsh metric since it ignores syntactic structure facts that explain how clusters of words move together.

| A-C | A-E | E-C |
|---|---|---|
| 0.1679 | 0.0846 | 0.1531 |

Table 2: Average Relative Alignment Length for pairs of Arabic (A), English (E) and Chinese (C)

Table 2 presents the ARAL scores for each language pair. These scores are computed over the *grow-diag-final* symmetrized alignment we use in our system (Koehn, 2004). $ARAL_{AC}$ is the highest and $ARAL_{AE}$ is the lowest. The average length of sentences is generally close among these languages (given the segmentation we use): Arabic is ~32 words, English is ~31 and Chinese is ~29. Arabic and English are much closer to each other than either to Chinese. This may be the result of Arabic tokenization and Chinese segmentation technologies which have been developed for translation into English. We address this issue in section 4.1. The ARAL scores agree with our assessment that English is closer to Arabic and to Chinese than Arabic is to Chinese. As a result, we believe it may serve as a good pivot language for translating Arabic to Chinese.

## 4 System Description

In this section, we describe the different systems we compare.

### 4.1 Data

Our data collection is the United Nations (UN) multilingual corpus, provided by the LDC [4] (catalog no. LDC2004E12). The UN corpus has in principle parallel sentences for Arabic, English and Chinese. However, the Arabic-English (A-E) data and Chinese-English (C-E) data sets were not in synch. The A-E data set has 3.2M lines while the C-E data set has 5.0M lines. We used the document ID provided in the data to match sentences from A-E against those in C-E to generate a three-way parallel corpus with 2.6M lines.

We tokenized the Arabic data in the Arabic Treebank scheme (Sadat and Habash, 2006). Chinese was segmented into words using a segmenter developed by Howard Johnson for the Portage Chinese-English MT system.[5] So a sentence consists of multiple words with spaces between them and each word is comprised of one or more characters. English was simply processed to split punctuation and "'s". The same preprocessing was used in all systems compared.

We are aware of two potentially biased aspects of our experimental setting. First, the Arabic and Chinese portions of our data collection, the UN corpus, are known to be generated from English originals. And secondly, the preprocessing techniques we used on Arabic and Chinese were developed for translation from these languages into English. These two aspects make English potentially more central to our experiments than if the data collection and preprocessing were done on Arabic and Chinese independent of English. Of course, it must be noted that the data bias is not unique to our work but rather a challenge for any bilingual corpus, in which translation is done from one language to another. Additionally, we can argue that the English bias in data and preprocessing does not only affect the Arabic-English and English-Chinese pipelines, but it also makes the Arabic and Chinese data potentially closer. Finally, given the expense involved in creating direct Arabic-Chinese parallel text and given the large amounts of Arabic-English and English-Chinese data, we think our results (with English bias) are still valid and interesting. That said, we leave the question of Arabic-Chinese optimization to future work.

### 4.2 Direct A-C MT System

In our baseline direct A-C system, we used the Arabic and Chinese portions of our parallel corpus to train a direct phrase-based MT system. We use GIZA++ (Och and Ney, 2003) for

[4] http://www.ldc.upenn.edu

[5] http://iit-iti.nrc-cnrc.gc.ca/projects-projets/portage_e.html

word alignment, and the Pharaoh system suite to build the phrase table and decode (Koehn, 2004). The Chinese language model (LM) used 200M words from the UN corpus segmented in a manner consistent with our training. The trigram LM was built using the SRILM toolkit (Stolcke, 2002).

### 4.3 Sentence Pivoting MT System

The sentence pivoting system (A-s-C) used English as an interface between two separate pharse-based MT systems: an Arabic-English direct system and an English-Chinese direct system. When translating Arabic to Chinese, the English top-1 output of the Arabic-English system was passed as input to the English-Chinese system. The English LM used to train the Arabic-English system is built from the counterpart of the Chinese data used to build the Chinese LM in our parallel corpus. We use 210M English words in total.

### 4.4 Phrase Pivoting MT System

The phrase pivoting system (A-p-C) extracts a new Arabic to Chinese phrase table using the Arabic-English phrase table and the English-Chinese phrase table. We consider a Chinese phrase a translation of an Arabic phrase only if some English phrase can bridge the two. We use the following formulae to compute the lexical and phrase probabilities in the new phrase table in a similar manner to Utiyama and Isahara (2007). Here, $\phi$ is the lexical probability and $p_w$ is the phrase probability.

$$\phi'(a \mid c) = \sum_e \phi(a \mid e)\phi(e \mid c)$$

$$\phi'(c \mid a) = \sum_e \phi(c \mid e)\phi(e \mid a)$$

$$p_w'(a \mid c) = \sum_e p_w(a \mid e)p_w(e \mid c)$$

$$p_w'(c \mid a) = \sum_e p_w(c \mid e)p_w(e \mid a)$$

The left hand side of the formulae represents the four required probabilities in a Pharaoh Arabic-Chinese phrase table.

## 5 Evaluation

For each of the direct system, the sentence-pivoting system and the phrase-pivoting system, we conduct four sets of experiments with different data sizes. Table 3 illustrates the training data size for each experiment. The training data is collected from the beginning of the same parallel corpus, so the larger training sets include the smaller ones.

| | Lines | Words (Arabic) |
|---|---|---|
| S | 32500 | 1 Million |
| M | 65000 | 2 Million |
| L | 130000 | 4 Million |
| XL | 260000 | 8 Million |

Table 3: Training Data Size

We use two other data sets (1K lines each) for tuning and testing. Each sentence in these sets has only one reference. Tuning and testing data sets are the same across all experiments and systems. In all our experiments, we decode using Pharaoh (Koehn, 2004) with a distortion limit of 4 and a maximum phrase length of 7. Tuning is done for each experimental condition using Och's Minimum Error Training (Och, 2003).

Note that for each set of experiments with the same data size, we draw Chinese, Arabic and English from the same chunk of three way parallel corpus. For example, in S size experiments, the two phrase tables used to build a new table in the phrase-pivoting approach are extracted respectively from the A-E and E-C systems built in the sentence-pivoting approach with size S corpora.

### 5.1 Direct System Results

Table 4 shows the results of the direct translation system A-C. It also includes the result for A-E and E-C direct translation. As expected, as we double the size of the data, the BLEU score (Papineni et al., 2002) increases. However, the rate of increase is not always consistent. In particular, the M and L conditions vary highly in A-E compared to A-C. This is odd especially given that we are comparing the same set of data from the three parallel corpora. We speculate that this may have to do with an oddity in that portion of the data set that may have a different quality than the rest. We see the effect of this drop in A-E in the next section. BLEU is measured on English case-insensitively. BLEU is measured on Chinese using segmented words not characters.

| | A-C | A-E | E-C |
|---|---|---|---|
| **S** | 11.17 | 21.89 | 19.29 |
| **M** | 13.43 (+20.2%) | 23.86 (+9.0%) | 20.85 (+8.1%) |
| **L** | 14.62 (+8.9%) | 24.86 (+4.2%) | 22.42 (+7.5%) |
| **XL** | 16.17 (+10.6%) | 27.96 (+12.5%) | 24.11 (+7.5%) |

Table 4: BLEU-4 scores comparing performance of direct translation of Arabic-Chinese (A-C), Arabic-English (A-E) and English-Chinese (E-C) for four training data sizes. The percentage increases are against the immediately lower data size.

## 5.2 Pivoting System Results

In Table 5, we present the results of the sentence pivoting system (A-s-C) and the phrase pivoting system (A-p-C). Under all conditions, A-s-C and A-p-C outperform A-C. A-p-C generally outperforms A-s-C except in the M data condition. The effect in the S conditions is bigger than the XL condition. In our best result (XL), we increase the BLEU score by over 1.12 points. Furthermore, the relative BLEU score increase from the L condition for A-p-C is 15.5% as opposed to A-C's 10.6%. The A-s-C relative increase from L to XL is 12.8%. This suggests that we are making better use of the available resources. The differences between A-s-C and A-C and between A-p-C and A-C are statistically significant at the 95% confidence level (Zhang et al., 2004). The differences between the two pivoting systems are not statistically significant. Examples from our best performing system are shown in Figure 2.

| | A-C | A-s-C | | A-p-C | |
|---|---|---|---|---|---|
| **S** | 11.17 | 12.24 | 9.6% | 13.12 | 17.5% |
| **M** | 13.43 | 14.10 | 5.0% | 13.75 | 2.4% |
| **L** | 14.62 | 14.96 | 2.3% | 14.97 | 2.4% |
| **XL** | 16.17 | 16.88 | 4.4% | 17.29 | 6.9% |

Table 5: Word-based BLEU-4 scores. A-C is direct translation. A-s-C is indirect translation through sentence pivoting and A-p-C is indirect translation through phrase pivoting. The percentages indicate relative improvement over A-C.

Our results are consistent with (Utiyama and Isahara, 2007) in that phrase-pivoting generally does better than sentence pivoting. However, we disagree with them in that, for us, direct translation is not the best system to use. We believe that this effect is caused by the combination of the very different languages we use. English is truly bridging between Arabic and Chinese in many linguistic dimensions. We think it's English's middle-ground-ness that makes these results possible.

| | | A-C | A-s-C | | A-p-C | |
|---|---|---|---|---|---|---|
| **BLEU-1** | **S** | 53.75 | 54.38 | 1.2% | 54.64 | 1.7% |
| | **M** | 56.65 | 57.00 | 0.6% | 55.88 | -1.4% |
| | **L** | 58.37 | 57.69 | -1.2% | 58.79 | 0.7% |
| | **XL** | **59.90** | **60.34** | **0.7%** | **60.28** | **0.6%** |
| **BLEU-4** | **S** | 21.32 | 21.80 | 2.3% | 22.88 | 7.3% |
| | **M** | 23.84 | 24.22 | 1.6% | 23.76 | -0.3% |
| | **L** | 24.98 | 25.14 | 0.6% | 25.87 | 3.6% |
| | **XL** | **25.95** | **27.11** | **4.5%** | **27.70** | **6.7%** |
| **BLEU-7** | **S** | 9.82 | 10.02 | 2.0% | 11.42 | 16.3% |
| | **M** | 11.56 | 11.84 | 2.4% | 11.64 | 0.7% |
| | **L** | 12.23 | 12.52 | 2.4% | 13.09 | 7.0% |
| | **XL** | **12.69** | **13.52** | **6.5%** | **14.57** | **14.8%** |

Table 6: Character-based BLEU scores for n-grams of maximum size 1, 4, and 7. The percentages are relative to the direct system.

In Table 6, we present additional scores using BLEU-1, BLEU-4 and BLEU-7 measured at the character level as opposed to the harsher measure at word level. Ignoring the odd behavior in M and L conditions, the sentence-pivot and phrase-pivot approaches improve over the direct translation baseline in terms of fluency (BLEU-7) and accuracy (BLEU-1). Under the small data condition, the phrase-pivot approach increases the BLEU-4 score three times the increase of the sentence-pivot approach. That ratio reduces to 1.5 times in the XL condition. The relative improvements of the pivoting systems over the direct system are small at BLEU-1 and much bigger at higher BLEU scores. This suggests that differences between the pivoting systems and the direct system are not in terms of lexical coverage but rather in terms of better reordering.

The lengths of the outputs of all the systems (direct and pivoting) are larger than the reference length which means no brevity penalty was applied in BLEU calculation. Also, no BLEU-gaming was done by OOV deletion: all OOV words were left in the output.

## 5.3 Error Analysis

We conducted an error analysis of our best performing system (Phrase Pivot XL) to understand what issues need to be addressed in the

future. We took a sample of 50 sentences restricted in length to be between 15 and 35 Chinese words. A Chinese native speaker compared our output to the reference translation and judged its quality in terms of two categories: syntax and lexical choice.

In terms of syntax, our judge identified all the occurrences of (a) subjects and verbs, (b) prepositional phrases and verbs and (c) modified nouns. Each case was judged as *acceptable* or *wrong*. Placing a verb before its subject, a preverbal prepositional phrase after its verb, or a modifier after the noun it modifies are all considered *wrong*. We correctly produce subject-verb order 73% of the time; and we produce nominal modification order correctly 64% of the time. Our biggest weakness in terms of syntax is prepositional phrase order. It is worth noting that the two phenomena we do better on are addressed in translation from Arabic to English, unlike prepositional phrase order which is where Chinese is different from both Arabic and English.

In terms of lexical choice our judge considered the translation quality of three classes of words: Nominals (nouns, pronouns, adjectives and adverbs), Verbs, and other particles (prepositions, conjunctions and quantifiers). An incorrectly translated or deleted word is considered *wrong*. We perform on nominals and particles at about the same level of 90%. Verbs are our biggest challenge with accuracy below 80%. The ratio of deleted words among all wrong words is rather high at about 30% (for nominals and for verbs). The detailed results of the error analysis are shown in Table 7.

Finally, there are 27 instances of Arabic Out-of-Vocabulary (OOV) words (1.93% of all words) that are not handled. Ten (37%) of these are proper nouns. The rest belong to mostly nouns and adjectives. Orthogonally, 19 (70%) of all OOV words belong to the genre of science reports, which is quite different from the data we train on. The OOVs include complex terms like السيبروفلوكساسين *Alsybrwflwksasyn* 'ciprofloxacin' and رجاجات مدارية *rjAjAt mdAryħ* '[chemical] orbital shakers'. Other less frequent OOV cases involve bad tokenization and less common morphological constructions.

| | | Total | Acceptable | Wrong |
|---|---|---|---|---|
| Syntax | Subj-Verb | 48 | 35(73%) | 13 (27%) |
| | Verb-PP | 46 | 17 (37%) | 29 (63%) |
| | Noun-Mod | 97 | 62 (64%) | 35 (36%) |
| Lexical Choice | Nominal | 408 | 368 (90%) | 40 (10%) |
| | Verb | 124 | 98 (79%) | 26 (21%) |
| | Particle | 116 | 106 (91%) | 10 (9%) |

Table 7: Results of human error analysis on a sample from the A-p-C system (XL)

## 6 Conclusion and Future Work

We presented a comparison of two approaches for Arabic-Chinese MT using English as a pivot language against direct MT. Our results show that using English as a pivot in either approach outperforms direct translation from Arabic to Chinese. We believe that this is a result of English being a sort of middle ground between Arabic and Chinese in terms of different linguistic features (in particular word order). Our best result is the phrase-pivot system which scores higher than direct translation by 1.1 BLEU points. An error analysis of our system shows that we successfully handle many complex Arabic-Chinese syntactic variations although there is a large space for improvement still.

In the future, we plan on exploring tighter coupling of Arabic and Chinese through comparing different methods of preprocessing Arabic for Arabic-Chinese MT, in a similar manner to Sadat and Habash (2006). We also plan to study how well these results carry on to different corpora (bilingual Arabic-English and English-Chinese) as opposed to the trilingual corpus used in this paper. We also plan to investigate whether our findings in Arabic-English-Chinese can be used for other different language triples.

## Acknowledgements

We would like to thank Roland Kuhn, George Foster and Howard Johnson of the National Research Council Canada for helpful advice and discussions and for providing us with the Chinese preprocessing tools.

Figure 1: An example highlighting Arabic-English-Chinese syntactic differences

يقرأ$_1$ الطالب$_2$ المجتهد$_3$ كتابا$_4$ عن$_5$ الصين$_6$ في$_7$ الصف$_8$ .
yqrÂ$_1$ AlTAlb$_2$ Almjthd$_3$ ktAbA$_4$ ςn$_5$ AlSyn$_6$ fy$_7$ AlSf$_8$ .
read$_1$ the-student$_2$ the-diligent$_3$ a-book$_4$ about$_5$ china$_6$ in$_7$ the-classroom$_8$ .

这$_1$位$_2$勤奋$_3$ 的$_4$学生$_5$ 在$_6$教室$_7$ 读$_8$ 一$_9$本$_{10}$ 关于$_{11}$ 中国$_{12}$ 的$_{13}$书$_{14}$。
this$_1$ quant$_2$ diligent$_3$ de$_4$ student$_5$ in$_6$ classroom$_7$ read$_8$ one$_9$ quant$_{10}$ about$_{11}$ china$_{12}$ de$_{13}$ book$_{14}$
Zhe$_1$ Wei$_2$ Qin Fen$_3$ De$_4$ Xue Sheng$_5$ Zai$_6$ Jiao Shi$_7$ Du$_8$ Yi$_9$ Ben$_{10}$ Guan Yu$_{11}$ Zhong Guo$_{12}$ De$_{13}$ Shu$_{14}$

*The diligent student is reading a book about China in the classroom.*

Figure 2: Examples of Arabic-Chinese MT output. English references and English glosses for Arabic and Chinese are provided to ease readability.

| | |
|---|---|
| Arabic | وبناء على ذلك ، فان هذه البيئة معرضة للفساد وانعدام الكفاءة الى حد بعيد .<br>and-building upon this , therefore this environment susceptible to-corruption and-lack qualification to extent big . |
| Eng-Ref | Consequently , this environment lends itself to significant degrees of corruption and inefficiency . |
| Chn-Ref | 因此,这种环境导致了高度腐败和效率低下。<br>Therefore, this kind environment caused have high-degree corruption and efficiency low. |
| Chn-Out | 因此,这种环境中的腐败和缺乏 效率在很大程度上 。<br>Therefore , this kind environment inside DE corruption and lack efficiency on big degree top. |
| Arabic | واذا لم ترد المعلومات المطلوبة في غضون 90 يوما اخرى ، يسقط الطلب .<br>and-if did-not arrive information requested in period 90 day other , lapse application . |
| Eng-Ref | If the requested information is not received within a further 90 days , the application will lapse. |
| Chn-Ref | 如果再过 90 天仍未收到所 要求的资料,则申请失效。<br>If again pass 90 day yet not received requested DE information , then application loose validity. |
| Chn-Out | 如果没有收到所 要求的资料 90 天内提供更多的要求。<br>If not receive requested DE information 90 day within provide more DE request. |
| Arabic | ... تيسير تبادل المعلومات والتشارك في الموارد بين الاجهزة الحكومية .<br>… facilitation exchanging the-information and-the-sharing in the-resources between the-agencies the-governmental . |
| Eng-Ref | … to facilitate the sharing of information and resources between government agencies . |
| Chn-Ref | …为各政府机构之间交流信息和资源提供便利 。<br>…for all government agencies among exchanging information and resource offer convenience. |
| Chn-Out | …旨在促进信息交流和分享资源政府间机构。<br>...purpose in facilitate information exchanging and sharing resource governments among agency . |
| Arabic | وينبغي للحكومات ان تنظر في استحداث تدابير مناسبة وفعالة للتقليل الى الحد الادنى من احتمالات الفساد .<br>and-should to-government that look in introducing measures appropriate and-effective to-reduce to extent least from possibilities the-corruption |
| Eng-Ref | Governments should consider introducing appropriate and effective measures to minimize the potential for corruption. |
| Chn-Ref | 各国政府应考虑采取适当的有效措施,最大 限度地 减少产生腐败的可能性 。<br>all countries governments should consider adopt appropriate DE effective methods , to-biggest-extent DE reduce producing corruption DE possibility. |
| Chn-Out | 各 国政府应考虑建立适当的有效措施,最大限度地减少腐败的可能性 。<br>all countries governments should consider build appropriate DE effective methods , to-biggest-extent DE reduce corruption DE possibility. |

## References


Yaser Al-Onaizan and Kishore Papineni. 2006 Distortion models for statistical machine translation. *In Proceedings of* Coling-ACL'06. Sydney, Australia.

Bogdan Babych, Anthony Hartley, and Serge Sharoff. 2007. Translating from under-resourced languages: comparing direct transfer against pivot translation. *In Proceedings of* MT Summit XI, Copenhagen, Denmark.

Tim Buckwalter. 2002. Buckwalter Arabic morphological analyzer version 1.0.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. *In Proceedings of HLT-NAACL'06.* New York, NY, USA.

Marine Carpuat and Dekai Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. *In Proceedings of* MT Summit XI, Copenhagen, Denmark.

Josep M. Crego and José B. Mariño. 2007. Syntax-enhanced n-gram-based SMT. *In Proceedings of* MT Summit XI, Copenhagen, Denmark.

Karim Filali and Jeff Bilmes. Leveraging Multiple Languages to Improve Statistical MT Word Alignments. *In Proceedings of* ASRU'05, Cancun, Mexico.

Nizar Habash. 2007. Syntactic preprocessing for statistical machine translation. *In Proceedings of* MT Summit XI, Copenhagen, Denmark.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In A. van den Bosch and A. Soudi, editors, Arabic Computational Morphology: Knowledge-based and Empirical Methods. Springer.

Hitoshi Isahara, Sadao Kurohashi, Jun'ichi Tsujii, Kiyotaka Uchimoto, Hiroshi Nakagawa, Hiroyuki Kaji, and Shun'ichi Kikuchi. 2007. Development of a Japanese-Chinese machine translation system. *In Proceedings of* MT Summit XI, Copenhagen, Denmark.

Philipp Koehn. 2004. Pharaoh: a Beam Search Decoder for Phrase-based Statistical Machine Translation Models. *In Proceedings of* AMTA'04, Washington, DC, USA.

Shankar Kumar, Franz Och, and Wolfgang Macherey. 2007. Improving word alignment with bridge languages. *In Proceedings of* EMNLP-CoNLL'07, Prague, Czech Republic.

Young-Suk Lee. 2004. Morphological Analysis for Statistical Machine Translation. *In Proceedings of* HLT-NAACL'04, Boston, MA, USA.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29 (1):19–52.

Franz Josef Och. 2003. Minimum Error Rate Training for Statistical Machine Translation. *In Proceedings of* ACL'03, Sapporo, Japan.

Fatiha Sadat and Nizar Habash. 2006. Combination of Arabic preprocessing schemes for statistical machine translation. *In Proceedings of* Coling-ACL'06. Sydney, Australia.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of* ACL'02, Philadelphia, PA, USA.

Charles Schafer & David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. *In Proceedings of* CoNLL'02, Taipei, Taiwan.

Micheal. Simard. 1999. Text translation alignment: Three languages are better than two. *In Proceedings of EMNLP-VLC'99, College Park, MD, USA.*

Michel Simard, Nicola Ueffing, Pierre Isabelle, and Roland Kuhn. 2007. Rule-based translation with statistical phrase-based post-editing. *In Proceedings* of the workshop on Statistical Machine Translation, ACL'07, Prague, Czech Republic.

Andreas Stolcke. 2002. SRILM - an Extensible Language Modeling Toolkit. I*n Proceedings of* ICSLP'02, Denver, CO, USA.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. *In Proceedings of* NAACL-HLT'07, Rochester, NY, USA.

Chao Wang, Michael Collins, and Philipp Koehn. 2007. Chinese syntactic reordering for statistical machine translation. *In Proceedings of* EMNLP-CoNLL'07, Prague, Czech Republic.

Dekai Wu. 1998. A Position Statement on Chinese Segmentation. *Presented at the Chinese Language Processing Workshop*. http://www.cs.ust.hk/~dekai/papers/segmentation.html.

Hua Wu and Haifeng Wang. 2007. Pivot language aproach for phrase-based statistical machine translation. *In Proceedings of* ACL'07, Prague, Czech Republic.

Yang Ye, Karl-Michael Schneider, and Steven Abney. 2007. Aspect marker generation in English-to-Chinese machine translation. *In Proceedings of* MT Summit XI, Copenhagen, Denmark.

Ying Zhang, Stephan Vogel and Alex Waibel, Interpreting Bleu/NIST scores: How much improvement do we need to have a better system?, *In Proceedings of* LREC'04, Lisbon, Portugal.