# Domain and Dialect Adaptation for Machine Translation into Egyptian Arabic

**Serena Jeblee**[1] **, Weston Feely**[1] **, Houda Bouamor**[2]
**Alon Lavie**[1]**, Nizar Habash**[3] **and Kemal Oflazer**[2]

[1]**Carnegie Mellon University**
{sjeblee, wfeely, alavie}@cs.cmu.edu
[2]**Carnegie Mellon University in Qatar**
hbouamor@qatar.cmu.edu, ko@cs.cmu.edu
[3]**New York University Abu Dhabi**
nizar.habash@nyu.edu

## Abstract

In this paper, we present a statistical machine translation system for English to Dialectal Arabic (DA), using Modern Standard Arabic (MSA) as a pivot. We create a core system to translate from English to MSA using a large bilingual parallel corpus. Then, we design two separate pathways for translation from MSA into DA: a two-step domain and dialect adaptation system and a one-step simultaneous domain and dialect adaptation system. Both variants of the adaptation systems are trained on a 100k sentence tri-parallel corpus of English, MSA, and Egyptian Arabic generated by a rule-based transformation. We test our systems on a held-out Egyptian Arabic test set from the 100k sentence corpus and we achieve our best performance using the two-step domain and dialect adaptation system with a BLEU score of 42.9.

## 1 Introduction

While MSA is the shared official language of culture, media and education in the Arab world, it is not the native language of any speakers of Arabic. Most native speakers are unable to produce sustained spontaneous discourse in MSA - they usually resort to repeated code-switching between their dialect and MSA (Abu-Melhim, 1991). Arabic speakers are quite aware of the contextual factors and the differences between their dialects and MSA, although they may not always be able to pinpoint exact linguistic differences. In the context of natural language processing (NLP), some Arabic dialects have started receiving increasing attention, particularly in the context of machine translation (Zbib et al., 2012; Salloum and Habash, 2013; Salloum et al., 2014; Al-Mannai et al., 2014) and in terms of data collection (Cotterell and Callison-Burch, 2014; Bouamor et al., 2014; Salama et al., 2014) and basic enabling technologies (Habash et al., 2012; Pasha et al., 2014). However, the focus is on a small number of iconic dialects, (e.g., Egyptian). The Egyptian media industry has traditionally played a dominant role in the Arab world, making the Egyptian dialect the most widely understood and used dialect. DA is now emerging as the language of informal communication online. DA differs phonologically, lexically, morphologically, and syntactically from MSA. And while MSA has an established standard orthography, the dialects do not: people write words reflecting their phonology and sometimes use roman script. Thus, MSA tools cannot effectively model DA; for instance, over one-third of Levantine verbs cannot be analyzed using an MSA morphological analyzer (Habash and Rambow, 2006). These differences make the direct use of MSA NLP tools and applications for handling dialects impractical.

In this work, we design an MT system for English to Egyptian Arabic translation by using MSA as an intermediary step. This includes different challenges from those faced when translating into English. Because MSA is the formal written variety of Arabic, there is an abundance of written data, including parallel corpora from sources like the United Nations and newspapers, as well as various treebanks. Using these resources, many researchers have created fairly reliable MSA translation systems. However, these systems are not designed to deal with the other Arabic variants.

Egyptian Arabic is much closer to MSA than it is to English, so one can get a system bet-

ter performance by translating first into MSA and then translating from MSA to Egyptian Arabic, which are far more similar. Our approach consists of a core MT system trained on a large amount of out-of-domain English-MSA parallel data, followed by an adaptation system. We design and implement two adaptation systems: a two-step system first adapts to in-domain MSA and then separately adapts from MSA to Egyptian Arabic, and a one-step system that adapts directly from out-of-domain MSA to in-domain Egyptian Arabic.

Our research contributions are summarized as follows:

(a) We build a machine translation system to translate into, rather than out of, dialectal Arabic (from English), using MSA as a pivot point.
(b) We apply a domain adaptation technique to improve the MSA results on our in-domain dataset.
(c) We automatically generate the Egyptian side of a 100k tri-parallel corpus covering MSA, English and Egyptian.
(d) We use this domain adaptation technique to adapt MSA into dialectal Arabic.

The remainder of this paper is structured as follows. We first review the main previous efforts for dealing with DA in NLP, in Section 2. In Section 3,we give a general description about using phrase-based MT as an adaptation system. Section 4 presents the dataset used in the different experiments. Our approach for translating English text into Egyptian Arabic is explained in Section 5. Section 6 presents our experimental setup and the results obtained. Then, we give an analysis of our system output in Section 7. Finally, we conclude and describe our future work in Section 8.

## 2 Related work

Machine translation (MT) for dialectal Arabic (DA) is quite challenging given the limited resources to build rule-based models or train statistical models for MT. While there has been a considerable amount of work in the context of standard Arabic NLP (Habash, 2010), DA is impoverished in terms of available tools and resources compared to MSA, e.g., there are few parallel DA-English corpora (Zbib et al., 2012; Bouamor et al., 2014). The majority of DA resources are for speech recognition, although more and more resources for machine translation and enabling technologies such as morphological analyzers are becoming available for specific dialects (Habash et al., 2012; Habash et al., 2013).

For Arabic and its dialects, several researchers have explored the idea of exploiting existing MSA rich resources to build tools for DA NLP. Different research work successfully translated DA to MSA as a bridge to translate to English (Sawaf, 2010; Salloum and Habash, 2013), or to enhance the performance of Arabic-based information retrieval systems (Shatnawi et al., 2012). Among the efforts on translation from DA to MSA, Abo Bakr et al. (2008) introduced a hybrid approach to transfer a sentence from Egyptian Arabic to MSA. Sajjad et al. (2013) used a dictionary of Egyptian/MSA words to transform Egyptian to MSA and showed improvement in the quality of machine translation. A similar but rule-based work was done by Mohamed et al. (2012). Boujelbane et al. (2013) and Hamdi et al. (2014) built a bilingual dictionary using explicit knowledge about the relation between Tunisian Arabic and MSA. These works are limited to a dictionary or rules that are not available for all dialects. Zbib et al. (2012) used crowdsourcing to translate sentences from Egyptian and Levantine into English, and thus built two bilingual corpora. The dialectal sentences were selected from a large corpus of Arabic web text. Then, they explored several methods for dialect/English MT. Their best Egyptian/English system was trained on dialect/English parallel data. They argued that differences in genre between MSA and DA make bridging through MSA of limited value. For this reason, while pivoting through MSA, it is important to consider the domain and add an additional step: domain adaptation.

The majority of previous efforts in DA MT has been focusing on translating from dialectal Arabic into other languages (mainly MSA or English). In contrast, in this work we focus on building a system to translate from English and MSA into DA. Furthermore, to the best of our knowledge, this is the first work in which we adapt the domain in addition to the dialect (Egyptian specifically).

## 3 Using Phrase-Based MT as an Adaptation System

For commercial use, MT output is usually post-edited by a human translator in order to fix the errors generated by the MT system. This is often faster and cheaper than having a human translate

the document from scratch. However, we can apply statistical phrase-based MT to create an automatic machine post-editor (what we refer to in this paper as an adaptation system) to improve the output of an MT system, and make it more closely resemble the references. Simard et al. (2007) used a phrase-based MT system as an automatic post-editor for the output of a commercial rule-based MT system, showing that it produced better results than both the rule-based system alone and a single pass phrase-based MT system. This technique is also useful for adapting to a specific domain or dataset. Isabelle et al. (2007) used a statistical MT system to automatically post-edit the output of a generic rule-based MT system, to avoid manually customizing a system dictionary and to reduce the amount of manual post-editing required.

For our adaptation systems, we build a core phrase-based MT system with a large amount of out-of-domain data, which allows us to have better coverage of the target language. For an adaptation system, we then build a second phrase-based MT system by translating the in-domain train, tune, and test sets through the core translation system, then using that data to build the second system. This system uses only in-domain data: parallel MT output from the core and the references. In this system, instead of learning to translate one language into another, the model learns to translate erroneous MT output into more fluent output of the same language, which more closely resembles the references.

In this work, we apply this technique for domain and dialect adaptation, treating Egyptian Arabic as the target language, and the MT-generated MSA as the erroneous MT output. We use this approach to adapt to the domain of the MSA data, and also to adapt to the Egyptian dialect. What we refer to as the "one-step" system is a core system plus one adaptation system, whereas the "two-step" system consists of the core plus two subsequent adaptation systems. We describe the systems in more detail in Section 5.

## 4 Data

For the core English to MSA system, we use the 5 million parallel sentences of English and MSA from NIST 2012 as the training set. The tuning set consists of 1,356 sentences from the NIST 2008 Open Machine Translation Evaluation (MT08) data (NIST Multimodal Information Group, 2010a), and the test set consists of 1,313 sentences from NIST MT09 (NIST Multimodal Information Group, 2010b).

We use a 5-gram MSA language model built using the SRILM toolkit (Stolcke, 2002) on 260 million words of MSA from the Arabic Gigaword (Parker et al., 2011). All our MSA parallel data and monolingual MSA language modeling data were tokenized with MADA v3.1 (Habash and Rambow, 2005) using the ATB (Arabic Treebank) tokenization scheme.

For the adaptation systems, we build a 100k tri-parallel corpus Egyptian-MSA-English corpus. The MSA and English parts are extracted from the NIST corpus distributed by the Linguistic Data Consortium. The Egyptian sentences are obtained automatically by extending Mohamed et al. (2012) method for generating Egyptian Arabic from morphologically disambiguated MSA sentences. This rule-based method relies on 103 transformation rules covering essentially nouns, verbs and pronouns as well as certain lexical items. For each MSA sentence, this method provides more than one possible candidate, in its original version, the Egyptian sentence kept was chosen randomly. We extend the selection algorithm by scoring the different sentences using a language model. For this, we use SRILM with modified Kneser-Ney smoothing to build a 5-gram language model. The model is trained on a corpus including articles extracted from the Egyptian version of Wikipedia[1] and the Egyptian side of the AOC corpus (Zaidan and Callison-Burch, 2011). We chose to include Egyptian Wikipedia for the formal level of sentences in it different from the regular DA written in blogs or microblogging websites (e.g., Twitter) and closer to the ones generated by our system.

We split this data into train, tune, and test sets of 98,027, 960, and 961 sentences respectively, after removing duplicates across sets. The MSA corpus was tokenized using MADA and the Egyptian Arabic data was tokenized with MADA-ARZ v0.4 (Habash et al., 2013), both using the ATB tokenization scheme, with alif/ya normalization.

## 5 System Design

Figure 1 shows a diagram of our three English to Egyptian Arabic MT systems: (1) the baseline MT system, (2) the one-step adaptation MT system, and (3) the two-step adaptation MT system. We describe each system below.

[1] http://arz.wikipedia.org/

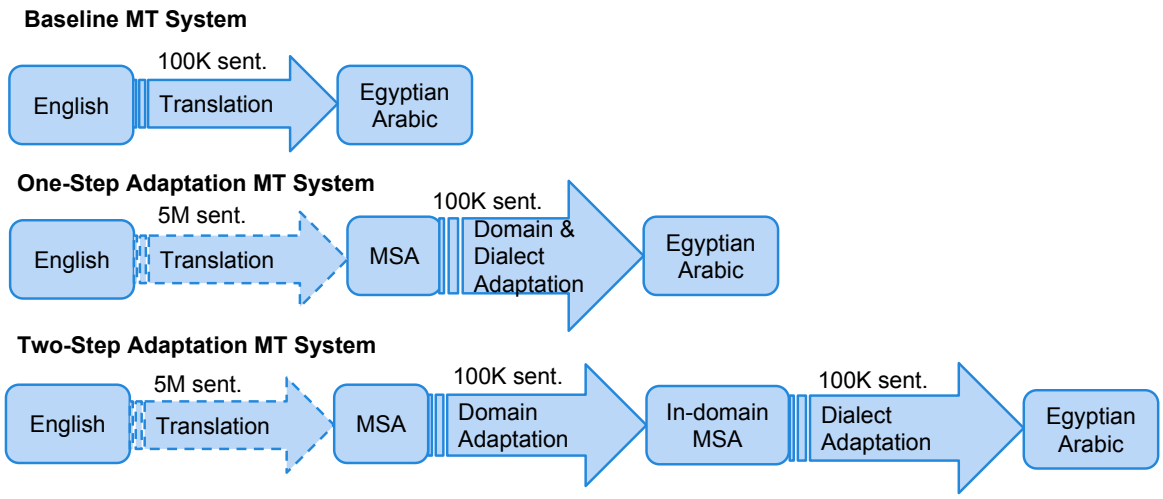


Figure 1: An overview of the different system architectures.

### Baseline System

Our baseline system is a single phrase-based English to Egyptian Arabic MT system, built using Moses (Koehn et al., 2007) on the 100k corpus described in Section 4. This system does not include any MSA data, nor does it have an adaptation system; it is a typical, one-pass MT system that translates English directly into Egyptian Arabic. We will show that using adaptation systems improves the results significantly.

### Core System

We base our systems on a core system built using Moses with the NIST data, a large amount of parallel English-MSA data from different sources than our in-domain data (the 100k dataset). Our core system is also built using Moses. We use this core system to translate the English side of our 100k train, tune, and test sets into MSA, the output of which we refer to as MSA'. This MSA' data is what we use as the source side for the adaptation systems.

### One-Step Adaptation System

To adapt to the domain and dialect of the 100k corpus, we first build a single adaptation system that translates the MSA' output of the core directly into Egyptian Arabic using the 100k corpus. The training data consists of parallel MSA' (the output of the core) and the Egyptian Arabic from the 100k dataset. With this system, we can take an English test set, translate it through the core to get MSA' output, which we can translate through the adaptation system to get Egyptian Arabic.

### Two-Step Adaptation System

We also build a two-step adaptation system that consists of two adaptation steps: one to adapt the MSA output of the core system to the domain of the MSA in the 100k corpus, and a second system to translate the MSA output of the domain adaptation system into Egyptian Arabic. We use the first adaptation system to translate the MSA' train, tune, and test sets (the output of the core, which is out-of-domain MSA), into in-domain MSA. This system is trained on the MSA' output parallel with the MSA references from the 100k dataset. We refer to the output of this system as MSA", because it has been translated from English into out-of-domain MSA (MSA'), and then from out-of-domain MSA to in-domain MSA.

The first adaptation system is used to translate the MSA' train, tune, and test sets into MSA". Then we use these MSA" sets with their parallel Egyptian Arabic from the 100k dataset to build the second adaptation system from in-domain MSA to Egyptian Arabic. We do not use the dialect transformation from (Mohamed et al., 2012) because it is designed to work with gold-standard annotation of the MSA text, which we do not have.

### System Variants

Since MSA and Egyptian are more similar to each other than they are to English, we tried several different reordering window sizes to find the optimal reordering distance for adapting MSA to Egyptian Arabic, including the typical reordering window of length 7, a smaller window of length 4, and no reordering at all. We found a reordering window

size of 7 to work best for all our systems, except for the one-step adaptation system, where no reordering produced the best result.

We also tested two different heuristics for symmetrizing the word alignments: grow-diag and grow-diag-final-and (Och and Ney, 2003). We found that using grow-diag as our symmetrization heuristic produced slightly better scores on the 100k datasets. For the baseline and adaptation systems we built 5-gram language models with KenLM (Heafield et al., 2013) using the target side of the training set, and for the core system we used the large MSA language model described in section 4. We use KenLM because it has been shown (Heafield, 2011) to be faster and use less memory than SRILM (Stolcke, 2002) and IRSTLM (Federico et al., 2008).

## 6 Evaluation and Results

For evaluation we use multeval (Clark et al., 2011) to calculate BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011), TER (Snover et al., 2006), and length of the test set for each system. We evaluate the core and adaptation systems on the MSA and Egyptian sides of the test set drawn from the 100k corpus, which we refer to as the 100k sets. The data used for evaluation is a genuine Egyptian Arabic generated from MSA, just like the data the systems were trained on. It is not practical to evaluate on naturally-generated Egyptian Arabic in this case because the domain of our datasets is very formal, since most of the text comes from news sources, and dialectal Arabic is generally used in informal situations.[2]

Below we report BLEU scores from our evaluation using tokenized and detokenized system output. We separate our results into the baseline system results, the results of the core, the results of the adaptation systems, and a comparison section. We specify scores of intermediate system output, such as MSA, as BLEU (A), and the scores of final system output as BLEU (B). For error analysis, we use METEOR X-ray (Denkowski and Lavie, 2011) to visualize the alignments of our system results with the references and each other.

For all MT systems we used grow-diag as our symmetrization heuristic. For each system, we report only the BLEU score of the best reordering window variant, which is specified in the caption below each table. The difference in scores between the different reordering window sizes (7, 4, and 0) we tried for the adaptation systems was not large (between 0 and 0.7 BLEU). In the following tables we present the best results for each adaptation system, which is a reordering window size of 7 for all systems, except for the phrase-based one-step domain and dialect adaptation system, which performs better with no reordering (0.2 BLEU better than a window of 7, 0.6 BLEU better than a window of 4), but these small differences in BLEU scores are within noise. The greatest difference in scores from the reordering windows was in the two-step systems domain adaptation step (MSA to MSA) on top of the phrase-based core, where a reordering window of 7 was 0.7 BLEU better than a window of 0.

### 6.1 Baseline System Results

| | **BLEU (B)** | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **100k EGY Tune** | 22.6 | 22.3 |
| **100k EGY Test** | 21.5 | 21.1 |

Table 1: Baseline results (English → EGY) with a reordering window size of 7.

The baseline system demonstrates the results of building a basic MT system directly from English to Egyptian Arabic. The goal of the core and adaptation systems is to achieve better scores than this initial approach.

### 6.2 Core System Results

In Table 2, we report BLEU scores for our core system on its own tuning set, NIST MT08, and NIST MT09 as a held-out MSA test set. We also report scores on the tune and test sets used to build our adaptation systems, both MSA and Egyptian Arabic. This is not the final system output, but rather these scores are for intermediate output only, which becomes the input for our adapatation systems.

We notice that unsurprisingly the core system performs much better on the 100k MSA test set than on the 100k Egyptian Arabic test set, which is to be expected because the core system is not trained on any Egyptian Arabic data. This shows the impact that the dialectal differences make on MT output. The results on the Egyptian test set here are the result of evaluating MSA output against Egyptian Arabic references.

[2]It is important to note that the Egyptian Arabic data we use is more MSA-like than typical Egyptian because it was generated directly from MSA.

| | BLEU (A) | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **NIST MT08 (Tune)** | 23.6 | 22.8 |
| **NIST MT09 (Test)** | 29.3 | 28.5 |
| **100k MSA Tune** | 39.8 | 39.3 |
| **100k MSA Test** | 39.4 | 39.0 |
| **100k EGY Tune** | 28.1 | 28.1 |
| **100k EGY Test** | 27.7 | 27.7 |

Table 2: Core system (English → MSA) results using a reordering window size of 7.

### 6.3 Adaptation System Results

The adaptation systems take as input the MSA output of the core and attempt to improve the scores on the Egyptian test set by adapting to the domain of the 100k dataset, as well as to Egyptian Arabic, in either one or two steps.

| | BLEU (B) | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **100k EGY Tune** | 40.8 | 40.5 |
| **100k EGY Test** | 40.3 | 40.1 |

Table 3: One-Step Adaptation system (MSA' → Egyptian Arabic) results using a reordering window size of 0.

Table 3 shows the results of the single adaptation system, which adapts directly from the MSA output of the core to Egyptian Arabic. These BLEU scores are already much better than the core systems performance on the same test sets, improving from 28.1 BLEU to 40.5 BLEU on the Egyptian Arabic tuning set (a difference of 12.4 BLEU) and improving from 22.7 BLEU to 40.1 BLEU on the Egyptian Arabic test set (a difference of 17.4 BLEU).

Tables 4 and 5 below illustrate the results of the first and second steps of the two-step adaptation system: Table 4 contains the results of the first domain adaptation step from out-of-domain MSA to in-domain MSA and Table 5 contains the results of the second dialect adaptation step from in-domain MSA to Egyptian Arabic.

An example of our system output for an English sentence is given in Table 6. Its METEOR X-ray alignment is illustrated in Figure 2.

### 6.4 System Comparisons on 100k Test Sets

In Table 7, we compare the results from the core and the results from the first step of the two-step adaptation system on the MSA test set and we see that adapting to the domain improves BLEU scores on MSA.

Since our goal is to improve the output for

| | BLEU (A) | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **100k MSA Tune** | 45.2 | 44.6 |
| **100k MSA Test** | 44.8 | 44.2 |
| **100k EGY Tune** | 32.2 | 32.2 |
| **100k EGY Test** | 32.0 | 32.0 |

Table 4: Domain Adaptation system (MSA' → MSA") for Two-Step Adaptation System Results using a reordering window size of 7.

| | BLEU (B) | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **100k EGY Tune** | 43.3 | 43.2 |
| **100k EGY Test** | 43.1 | 42.9 |

Table 5: Dialect Adaptation system (MSA" → Egyptian) for Two-Step Adaptation System Results using a reordering window size of 7.

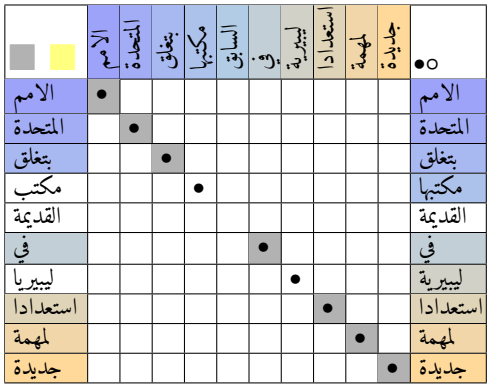


Figure 2: METEOR X-ray alignment of the sentence in table 6. The left side is the output of the one-step system, the right side is the output of the two-step system, and the top is the reference. The shaded cells represent matches between the reference and the one-step system, and the dots represent matches between the reference and the two-step system.

[1]One-Step System: Core + Domain and Dialect Adaptation (MSA' → EGY)

[2]Two Step Adaptation System (Step 1): Core + Domain Adaptation (MSA' → MSA")

[3]Two Step Adaptation System (Step 2): Core + Domain Adaptation (MSA' → MSA") + Dialect Adaptation (MSA" → EGY)

| | |
|---|---|
| **English** | UN closes old office in Liberia in preparation for new mission |
| **Egyptian Reference** | الامم المتحدة بتغلق مكتبها السابق في ليبيرية استعدادا لمهمة جديدة<br>*AAlAmm AAlmtHdp btglq mktbhA AAlsAbq fy lybyryp AAstEdAdA lmhmp jdydp* |
| **1-Step System** | الامم المتحدة بتغلق مكتب القديمة في ليبيريا استعدادا لمهمة جديدة<br>*AAlAmm AAlmtHdp btglq mktb AAlqdymp fy lybyryA AAstEdAdA lmhmp jdydp* |
| **2-Step System (step2)** | الامم المتحدة بتغلق مكتبها القديمة في ليبيرية استعدادا لمهمة جديدة<br>*AAlAmm AAlmtHdp btglq mktbhA AAlqdymp fy lybyryp AAstEdAdA lmhmp jdydp* |

Table 6: An example of system output from the Egyptian test set.

| | **BLEU (A)** | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **Core** (English → MSA') | 39.4 | 39.0 |
| **Core + Domain Adaptation** (MSA' → MSA") | 44.8 | 44.2 |

Table 7: Comparison of results on 100k MSA test set.

| | **BLEU (A/B)** | |
|---|---|---|
| | **Tokenized** | **Detokenized** |
| **Baseline** (English → EGY) | 21.5 (B) | 21.1 |
| **Core** (English → $MSA'$) | 27.7 (A) | 27.7 |
| **One-Step Adaptation System** [1] | 40.3 (B) | 40.1 |
| **Two-Step Adaptation System (Step 1)**[2] | 32.0 (A) | 32.0 |
| **Two-Step Adaptation System (Step 2)**[3] | 43.1 (B) | 42.9 |

Table 8: Comparison of results on 100k EGY test data.

| | **BLEU (B)** | **METEOR** | **TER** | **Length** |
|---|---|---|---|---|
| **Baseline System** | 21.1 | 38.5 | 66.1 | 102.7 |
| **One-Step System** | 40.1 | 53.4 | 51.3 | 100.0 |
| **Two-Step System: Step 2 (Dialect)** | 42.9 | 55.2 | 50.4 | 100.1 |

Table 9: Detokenized BLEU, METEOR, TER, and length scores for the best system results.

Egyptian Arabic, we examine the improvement of scores through different steps of the system in Table 8. These scores are all based on the same Egyptian Arabic references, even though some of the systems are designed to produce MSA output. It is important to note that although the first step of the two-step adaptation system (domain adaptation) is still producing MSA output, it performs better on the Egyptian test set than the out-of-domain MSA core. The domain adaptation system built on top of the core performs better than the core alone on the 100k corpus MSA test set (+5.2 BLEU), as well as the 100k corpus Egyptian Arabic test set (+4.3 BLEU). The best score we achieve on the 100k corpus MSA test set is 44.2 BLEU, from the core plus the domain adaptation system.

Table 9 shows the other detokenized scores from multeval (Clark et al., 2011) from the final output on the EGY test set from each system, and Table 10 shows BLEU-1 through BLEU-4 scores on the same detokenized results, which shows an improvement at different n-gram levels in unigram coverage from the baseline system to the adaptation systems.

Overall, the two-step adaptation system built on top of the core performs 15.2 BLEU better than the core alone on the 100k corpus Egyptian Arabic test set and the one-step adaptation system performs 12.4 BLEU better than the core on the same test set. The best score on the 100k EGY test set is from the two-step adaptation system with 42.9 BLEU, which outperforms the one-step adaptation system by 2.8 BLEU points. We consider possible causes of these results in section 7.

| | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
|---|---|---|---|---|
| **Baseline System** | 53.4 | 26.6 | 15.3 | 9.1 |
| **One-Step System** | 64.3 | 43.5 | 33.5 | 27.1 |
| **Two-Step System: Step 2 (Dialect)** | 65.2 | 46.0 | 36.8 | 30.7 |

Table 10: Detokenized BLEU (B) scores on the 100k EGY test set at different n-gram levels.

| | |
|---|---|
| **English** | US , Indonesia commit to closer trade , investment ties |
| **Egyptian Reference** | الولايات المتحدة واندونيسيا بيلتزموا بعلاقات تجارية واستثمارية اوثق<br>*AAlwlAyAt AAlmtHdp wAndwnysyA byltzmwA bElAqAt tjAryp wAstvmAryp AAwvq* |
| **Baseline output** | +نا ، فان اندونيسيا بتّعهد بتوثيق العلاقات التجارية والاستثمارية<br>*+nA , fAn AAndwnysyA bt Ehd btwvyq AAlElAqAt AAltjAryp wAlAstvmAryp* |

Table 11: An example of a Baseline system output sentence with no word matches.

| | |
|---|---|
| **English** | Pakistan sends envoys to Arab countries |
| **Egyptian Reference** | باكستان بترسل مبعوثين الي الدول العربية<br>*bAkstAn btrsl mbEwvyn AAly AAldwl AAlErbyp* |
| **One-Step System** | باكستان بيرسل عنثيس الي الدول العربية<br>*bAkstAn byrsl Envys AAly AAldwl AAlErbyp* |
| **Two-Step System (Step 2)** | باكستان بترسل عنثيس الي الدول العربية<br>*bAkstAn btrsl Envys AAly AAldwl AAlErbyp* |

Table 12: An example of system output from the Egyptian test set.

## 7 Error Analysis

In some of the output sentences, there are no exact matches and the sentence gets a score of 0, such as in the example from the Baseline system output in Table 11. But there are actually four words in the output that are present in the reference, but they have different clitics attached to them. The third word in the reference, واندونيسيا/*wAndwnysyA* "and Indonesia", is present in the output as just اندونيسيا/*AndwnysyA* "Indonesia". The same is true of the fifth, sixth, and seventh words in the reference: بعلاقات/*bElAqAt* "with relationships" is العلاقات/*AlElAqAt* "the relationships" in the output, تجارية/*tjAryp* "commercial" is التجارية/*AltjAryp* "commercial(definite)", and واستثمارية/*wAstvmAryp* "and investment" is والاستثمارية/*wAlAstvmAryp* "and the investment". In tokenized output the base words would be matched because the clitics would be separate. This is one of the drawbacks of evaluating on detokenized data.

Table 12 and Figure 3 show the output for a sentence from the Egyptian test set from the two different adaptation systems. In Figure 3, the results from the one-step and two-step adaptation systems are almost the same, except that the two-step adaptation system (which scored 2.8 BLEU higher than the one-step system overall) has one more word correct (the second word). This word is actually the same verb, but the two-step adaptation system has produced the correct conjugation of the verb (3rd person feminine), while the one-step system produced the wrong conjugation (3rd person masculine). In adapting to the domain first, the system seems to produce better subject-verb agreement.

In Table 6 and Figure 2 in Section 6.3, the transliteration of "Liberia" in the output of the two-step system matches the reference. The one-step system produces a different transliteration which is also valid, but is not the same one the reference uses. It also produces the correct object clitic (مكتبها/*mktbhA* "its office" vs. مكتب/*mktb* "office"). The output of the two-step system more consistently matches the reference in transliteration, subject-verb agreement, and clitic attachment.

In general the output of the two-step adaptation system appears to be in the correct order more often than the output of the one-step adaptation sys-

| | |
|---|---|
| **English** | man stabs nine at moscow synagogue |
| **Egyptian Reference** | شاب بيطعن ٩ في كنيس يهودي في موسكوا<br>*$Ab byTEnn 9 fy knys yhwdy fy mwskwA* |
| **One-Step System** | راجل طعن تسعه في كنيس يهودي في موسكوا<br>*rAjl TEn tsEh fy knys yhwdy fy mwskwA* |

Table 13: Comparison of reference and system output.

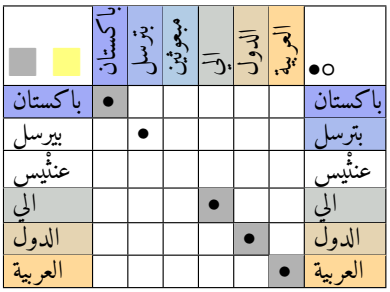


Figure 3: A comparison of the output of the one-step domain and dialect adaptation system (left column) and the two-step domain and dialect adaptation system (right column), both built on top of the phrase-based core. The top is the reference sentence.

tem, perhaps because we used a reordering window of 7 for the two-step system, whereas we used a window of 0 for the one step system. Additionally, the two-step system allows two passes of reordering, one in each step. Each step of the system produces a decrease in the fragmentation of the output: the output of the core on the Egyptian test set gets a fragmentation penalty of 0.204, the one-step system gets a fragmentation penalty of 0.159, and the two step system gets 0.189 for the first step (domain) and 0.139 for the second step (dialect). Since the output of the two-step system is less fragmented, there are longer sequences of words that are in the correct order.

Additionally, the one-step system misses more words, especially at the beginning of a sentence. There are many ways to introduce a sentence in Arabic, some of which correspond to the same English phrase. While the model will generate the most probable one, there may be several acceptable choices, and the reference may have a different one. For instance, in Table 13, the word "man" is translated as شاب/*$Ab* in the reference, and راجل/*rAjl* in the output of the one-step adaptation system. This word is penalized for not matching the reference, even though both are reasonable translations of "man". This problem could be helped by synonym matching in the evaluation metrics, which is not currently available for Arabic.

## 8 Conclusion and Future Work

We have shown that we can leverage a large amount of out-of-domain MSA data and a domain adaptation system to achieve better performance on an in-domain test set. We apply the same technique to translating Arabic dialects, by adapting from MSA to the Egyptian Arabic dialect as we would adapt between domains of the same language. Our results also show that when adapting to the domain, first by translating to MSA as an intermediary step and then adapting to the dialect, we can improve performance even more. Our results also show the importance of consistent and appropriate tokenization of the data. The tri-parallel corpus of English, MSA, and Egyptian gave us a unique opportunity to create this kind of system, as parallel data for Arabic dialects is hard to come by. However, this data is artificial Egyptian, not natural generated dialectal Arabic. In the future we hope to test our domain and dialect adaptation MT systems on more authentic Egyptian Arabic data sets and to be able to apply this technique to other Arabic dialects.

## Acknowledgements

This publication was made possible by grant NPRP-09-1140-1-177 from the Qatar National Research Fund (a member of the Qatar Foundation) and by computing resources provided by the NSF-sponsored XSEDE program under grant TG-CCR110017. The statements made herein are solely the responsibility of the authors. We thank the reviewers for their comments. Nizar Habash performed most of his contribution to this paper while he was at the Center for Computational Learning Systems at Columbia University.

## References


Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. In *Proceedings of the 6th International Conference on Informatics and Systems (INFOS2008)*. Cairo University.

Abdel-Rahman Abu-Melhim. 1991. Code-switching and Linguistic Accommodation in Arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*, volume 80, pages 231–250. John Benjamins Publishing.

Kamla Al-Mannai, Hassan Sajjad, Alaa Khader, Fahad Al Obaidli, Preslav Nakov, and Stephan Vogel. 2014. Unsupervised Word Segmentation Improves Dialectal Arabic to English Machine Translation. In *Proceedings of EMNLP Workshop on Arabic Natural Language Processing*, Doha, Qatar.

Houda Bouamor, Nizar Habash, and Kemal Oflazer. 2014. A Multidialectal Parallel Corpus of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1240–1245, Reykjavik, Iceland.

Rahma Boujelbane, Mariem Ellouze Khemekhem, and Lamia Hadrich Belguith. 2013. Mapping Rules for Building a Tunisian Dialect Lexicon and Generating Corpora. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 419–428, Nagoya, Japan.

Jonathan Clark, Chris Dyer, Alon Lavie, and Noah Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the Association for Computational Linguistics (ACL)*, Portland, Oregon.

Ryan Cotterell and Chris Callison-Burch. 2014. A Multi-Dialect, Multi-Genre Corpus of Informal Written Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 241–245, Reykjavik, Iceland.

Michael Denkowski and Alon Lavie. 2011. 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*, pages 1–4, Edinburgh, Scotland.

Marcello Federico, Nicola Bertoldi, and Mauro Cettolo. 2008. IRSTLM: an open source toolkit for handling large scale language models. In *INTERSPEECH*, pages 1618–1621.

Nizar Habash and Owen Rambow. 2005. Arabic Tokenization, Part-of-speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the Association for Computational Linguistics*, Ann Arbor, Michigan.

Nizar Habash and Owen Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 681–688, Sydney, Australia.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada.

Nizar Habash, Ryan Roth, Owen Rambow, Ramy Eskander, and Nadi Tomeh. 2013. Morphological Analysis and Disambiguation for Dialectal Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 426–432, Atlanta, Georgia.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*, volume 3. Morgan & Claypool Publishers.

Ahmed Hamdi, Nuria Gala, and Alexis Nasr. 2014. Automatically Building a Tunisian Lexicon for Deverbal Nouns. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 95–102, Dublin, Ireland.

Kenneth Heafield, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modfied Kneser-Ney Language Model Estimation. In *In Proceedings of the Association for Computational Linguistics*, Sofia, Bulgaria.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom, July.

Pierre Isabelle, Cyril Goutte, and Michel Simard. 2007. Domain Adaptation of MT Systems through Automatic Post-Editing. In *Proceedings of MT Summit XI*, pages 255–261, Copenhagen, Denmark.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Christopher Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Christopher Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 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*, pages 177–180, Prague, Czech Republic.

Emad Mohamed, Behrang Mohit, and Kemal Oflazer. 2012. Transforming Standard Arabic to Colloquial Arabic. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 176–180, Jeju Island, Korea.

NIST Multimodal Information Group. 2010a. NIST 2008 Open Machine Translation (OpenMT) Evaluation LDC2010T21. Web Download.

NIST Multimodal Information Group. 2010b. NIST 2009 Open Machine Translation (OpenMT) Evaluation LDC2010T23. Web Download.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. In *Computational Linguistics*, pages 19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the Association for Computational Linguistics*, Philadelphia, Pennsylvania.

Robert Parker, David Graff, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2011. Arabic Gigaword Fifth Edition LDC2011T11. Web Download.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1094–1101, Reykjavik, Iceland.

Hassan Sajjad, Kareem Darwish, and Yonatan Belinkov. 2013. Translating Dialectal Arabic to English. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Sofia, Bulgaria.

Ahmed Salama, Houda Bouamor, Behrang Mohit, and Kemal Oflazer. 2014. YouDACC: the Youtube Dialectal Arabic Comment Corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1246–1251, Reykjavik, Iceland.

Wael Salloum and Nizar Habash. 2013. Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 348–358, Atlanta, Georgia.

Wael Salloum, Heba Elfardy, Linda Alamir-Salloum, Nizar Habash, and Mona Diab. 2014. Sentence Level Dialect Identification for Machine Translation System Selection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 772–778, Baltimore, Maryland.

Hassan Sawaf. 2010. Arabic Dialect Handling in Hybrid Machine Translation. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 10)*, Denver, Colorado.

Mohammed Q Shatnawi, Muneer Bani Yassein, and Reem Mahafza. 2012. A Framework for Retrieving Arabic Documents Based on Queries Written in Arabic Slang Language. *Journal of Information Science*, 38(4):350–365.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical Phrase-Based Post-Editing. In *Proceedings of NAACL-HLT-2007 Human Language Technology: the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 508–515, Rochester, NY.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of Association for Machine Translation in the Americas*, Cambridge, Massachusetts.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing, vol. 2*, pages 901–904, Denver, CO, USA.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine Translation for Arabic Dialects. In *Proceedings of North American Chapter of the Association for Computational Linguistics*, Montreal, Canada.