Low Resourced Machine Translation via Morpho-syntactic Modeling: The Case of Dialectal Arabic

Alexander Erdmann ae1541@nyu.edu
Computational Approaches to Modeling Language (CAMeL) Lab,
New York University Abu Dhabi
Department of Linguistics, Ohio State University

Nizar Habash nizar.habash@nyu.edu
Computational Approaches to Modeling Language (CAMeL) Lab,
New York University Abu Dhabi

Dima Taji dima.taji@nyu.edu
Computational Approaches to Modeling Language (CAMeL) Lab,
New York University Abu Dhabi

Houda Bouamor hbouamor@cmu.edu
Department of Computer Science, Carnegie Mellon University Qatar

Abstract

We present the second ever evaluated Arabic dialect-to-dialect machine translation effort, and the first to leverage external resources beyond a small parallel corpus. The subject has not previously received serious attention due to lack of naturally occurring parallel data; yet its importance is evidenced by dialectal Arabic's wide usage and breadth of inter-dialect variation, comparable to that of Romance languages. Our results suggest that modeling morphology and syntax significantly improves dialect-to-dialect translation, though optimizing such data-sparse models requires consideration of the linguistic differences between dialects and the nature of available data and resources. On a single-reference blind test set where untranslated input scores 6.5 BLEU and a model trained only on parallel data reaches 14.6, pivot techniques and morpho-syntactic modeling significantly improve performance to 17.5.

1 Introduction

Arabic is widely spoken and highly diglossic, with Modern Standard Arabic (MSA) representing the high register shared across the Arab World in educated circles. In contrast, the many spoken dialectal Arabic varieties (DA) are somewhat if not entirely mutually unintelligible, e.g., Moroccan and Kuwaiti. Chiang et al. (2006) compare the linguistic variation among Arabic dialects to that among Romance languages, indicating the need for machine translation (MT)

between these dialects. However, while much MT research has been devoted to translating between Romance languages (Corbí Bellot et al., 2005; Armentano-Oller et al., 2006; Koehn et al., 2009), we are aware of only one work on Arabic DA-to-DA MT (Meftouh et al., 2015). It deals mainly with Maghrebi dialects and utilizes only a small parallel corpus.[1] This work focuses on the Egyptian and Levantine dialects, leveraging various available resources such as a morphological analyzer and additional monolingual and multilingual data.[2] Compared to other dialects, Egyptian and Levantine's wider range of available data/resources allows us to evaluate more MT approaches using different combinations of these data/resources. Thus, in future work on DA pairs which may not have the same data/resources, we can tailor MT systems based on this paper's findings.

The main challenge in developing DA-to-DA MT systems is the lack of data. While many Romance languages are official languages with written standards, naturally occurring in parallel corpora like the European Parliament (Koehn, 2005), DA has no official status and was rarely written until the advent of social media.[3] The recent release of the first parallel multi-dialectal corpora (Bouamor et al., 2014; Meftouh et al., 2015) has enabled seminal, albeit low-resource MT experiments. We present some shortcomings of these corpora and introduce an in-house, under-development corpus. Then we explore different means of leveraging external resources, e.g., Egyptian-to-English and Levantine-to-English data and an Egyptian tokenizer and morphological analyzer (Habash et al., 2012b; Maamouri et al., 2014; Pasha et al., 2014). We conduct experiments in a range of data-sparse settings and show the effect of morpho-syntactic features on the DA-to-DA MT performance. Our approach can be extended to other DA pairs and other closely related languages and dialects (Tyers et al., 2017).

2 Related Work

An increasing amount of research has been conducted on dialectal Arabic NLP; however, most dialectal MT efforts translate from DA to MSA or English. The only other DA-to-DA work we are aware of focuses on manipulating language model smoothing parameters to optimize data sparse MT performance (Meftouh et al., 2015).

2.1 Dialectal Arabic Machine Translation

While only Meftouh et al. (2015) have evaluated DA-to-DA MT, many others have addressed MT between DA and other languages. Zbib et al. (2012) attempted to translate from Egyptian and Levantine to English and found that pivoting through better resourced MSA was not useful due to register and domain differences. MSA, the higher register, is rarely used to discuss the day-to-day matters frequently treated with DA, causing a domain mismatch. However, several approaches have since presented alternative results (Sawaf, 2010; Salloum and Habash, 2011, 2012; Sajjad et al., 2013; Durrani et al., 2014). These use rule-based or hybrid methods to identify mappings from DA to MSA before translating to a target language (usually English). Additionally, Tachicart and Bouzoubaa (2014) report results on adapting an approach designed for MSA to Moroccan translation to translate in the inverse direction (Moroccan to MSA).

2.2 Dialectal Arabic Data

Several newly developed corpora have facilitated the recent surge in dialectal NLP work.

[1] Maghrebi dialects are those spoken in Morocco, Algeria, Tunisia and Libya.

[2] Levantine covers the dialects spoken in Lebanon, Syria, Palestine and Jordan.

[3] Recently, two parallel Arabic translations were created for 12,000 sentences from the European Parliamentary proceedings, but both are in MSA (Habash et al., 2017).

The DARPA BOLT (Broad Operational Language Translation) project sponsored the creation of a large number of resources,[4] including a sizeable data set of DA sentences paired with their English translations. This data set consists of 2.2 million words of Egyptian and 1.5 million words of Levantine which were harvested from SMS messages and online sources like weblogs before being translated.

As for monolingual corpora, Zaidan and Callison-Burch (2011)'s Arabic Online Commentary (AOC) corpus contains 52 million words of mixed MSA and DA from news articles and readers' comments. Cotterell and Callison-Burch (2014) add modest amounts of Twitter data to this corpus, though we find the domain difference harmful for language modeling and drop it in our experiments. Khalifa et al. (2016)'s GUMAR corpus contains over 100 million words of Gulf Arabic and a smattering of other dialects, all taken from internet novels, a genre of long conversational novels shared anonymously on online forums popular among female teenagers. Other monolingual DA corpora like Tunisiya (McNeil and Faiza, 2011), the Curras corpus of Palestinian-Levantine (Jarrar et al., 2014), and those corpora presented in Al-Shargi et al. (2016), focus on different dialects or are too small to be relevant to Egyptian-to-Levantine MT.

As for DA-to-DA data, Bouamor et al. (2014) present the first corpus with 2,000 7-way parallel sentences of Egyptian, Tunisian, three Levantine dialects (Syrian, Jordanian, Palestinian), MSA, and English, all translated from Egyptian sentences harvested from the web. The authors concede that many Levantine sentences seem to be influenced by the Egyptian, likely because translators were primed with Egyptian expressions they might understand, but would not produce naturally. The same concern applies to the 6,400 sentence, 6-way parallel PADIC corpus used in Meftouh et al. (2015), as all translations were derived from DA or MSA. When developing the 12,000 sentence multi-dialectal corpus used in our experiments, we avoided such priming effects by asking translators to produce translations starting from English sentences taken from the Basic Travel Expressions Corpus (BTEC) (Takezawa et al., 2002).

Other relevant resources include AVIA,[5] a small but rich multi-dialect reference grammar with contextual examples, and Tharwa (Diab et al., 2014), a 4-way English, MSA, Egyptian, Levantine lexicon with rich linguistic annotation.

2.3 Pivot Machine Translation

Pivoting is an MT technique used to combat data sparsity when more source-to-pivot and pivot-to-target data is available than source-to-target parallel data (Muraki, 1987; Hajič et al., 2004; Wu and Wang, 2007; Habash and Hu, 2009). In this work, we use a specific form of pivoting: phrase pivoting. This involves aligning source-to-pivot and pivot-to-target data, extracting pairs of phrases into two phrase tables, then combining them into a single source-to-target phrase table based on shared pivot phrases (Utiyama and Isahara, 2007).

Our work is similar to El Kholy et al. (2013), who use English to translate from Persian to Arabic via phrase pivoting. They introduce connectivity strength constraints to weight learned Persian-to-Arabic phrase-pairs in the table by considering how well each pair can be aligned through an English pivot phrase (discussed further in Sections 4 and 5). In follow-up work, El Kholy and Habash (2015) add morphological constraints for translating related, morphologically rich languages Arabic and Hebrew, via morphologically-poor English. These constraints help preserve fine grained morphological distinctions like gender agreement which cannot otherwise be accurately translated via a morphologically poor pivot that does not make such distinctions, i.e., English.

[4] Pointers to the Linguistic Data Consortium's BOLT resources can be found here: `https://www.ldc.upenn.edu/collaborations/current-projects/bolt`.

[5] `http://www.umventures.org/technologies/arabic-variant-identification-aid-avia`

3 Data Preparation

All data used in our experiments comes from sources mentioned in Section 2.2. As displayed in Table 1, we split our 12,000 sentence BTEC parallel corpus into training, tuning, dev, and blind test sets, which are constant across all experiments. Also, in some experiments, we use additional monolingual and pivot data from AOC and the BOLT corpus, respectively.

Data Set	Dialect	Description	Size
BTEC-train	Egy–Lev	Parallel	8,000
BTEC-tune	Egy–Lev	Parallel	500
BTEC-dev	Egy–Lev	Parallel	1,500
BTEC-test	Egy–Lev	Parallel	2,000
BOLT-egy	Egy–Eng	Pivot	410,000
BOLT-lev	Eng–Lev	Pivot	180,000
AOC-egy	Egy	Monolingual	9,000
AOC-lev	Lev	Monolingual	5,000

Table 1: Data used in all experiments. Size reported in number of sentences.

Similar to MSA, DA is morphologically and syntactically rich, posing several challenges for MT systems. To be able to leverage morpho-syntactic features, we ran our Egyptian and Levantine data through MADAMIRA (Pasha et al., 2014), an Arabic morphological analyzer and disambiguator trained for MSA (MADAMIRA-MSA) and Egyptian (MADAMIRA-EGY). Unfortunately, the Levantine version of MADAMIRA is still under development (Eskander et al., 2016), so we use MADAMIRA-EGY to process both our Egyptian and Levantine corpora. Jarrar et al. (2014) and Khalifa et al. (2016) show that using MADAMIRA-EGY to process non-Egyptian DA data yields better results than MADAMIRA-MSA. To minimize the analyzer's bias towards Egyptian when processing Levantine data, we do not allow it to make orthographic changes. This limits the effects of misanalyzing many Levantine words, such as هالحظ *hAlHĎ* 'this luck', which can be incorrectly Epgyptian*ized* as حألحظ *HÂlHĎ* – the Egyptian future particle +ح *H+* together with an MSA verb ألحظ *ÂlHĎ* 'I perceive'.[6]

A number of tokenization and segmentation schemes are available for Arabic (Habash, 2010). Some separate only punctuation and digits. Others, such as ATB and D3, separate different sets of clitics from the base word. Whereas D3 segments all clitics, ATB leaves attached the definite article, ال*Al*. The optimal segmentation for our task is D3 (Sadat and Habash, 2006), as the aggressive tokenization mitigates for data sparsity. Typically, these tokenization schemes involve orthographic rewrite rules to ensure that the base word matches its non-cliticized form to minimize sparsity (El Kholy and Habash, 2012). Such rules depend on the morphological template of the word and the clitics attached to it. For a word such as حيكتبوها *HyktbwhA* 'and they will write it', the basic D3 tokenization is *H+ yktbwA +hA*. The extra *A* is added to the base word to minimize sparsity as this is how it would appear if no suffix had been appended.[7]

Since we do not have ideal tools for processing (tokenizing and detokenizing) Levantine, we opt for a stricter surface-word-oriented segmentation that guarantees recovering the form by simple concatenation when detokenizing. Thus, for حيكتبوها *HyktbwhA* 'and they will write it', the desired D3 segmentation is *H+ yktbw +hA*. This may increase data sparsity slightly, but more importantly, as mentioned previously, this limits the extent to which words can be

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

[7] In all of the work presented in this paper we apply ا/ى *Alif/Ya* normalization (El Kholy and Habash, 2012).

			Required Data		
Model	**BLEU**	**Out-of-vocabulary**	**Parallel**	**Monolingual**	**Pivot**
NO-TRANSLATION	6.48	N/A			
DIRECT	15.44	4.6	X		
SYNTHETIC	16.75	0.8	X	X	
PHRASE PIVOT	6.77	1.4			X
DIR+PP	**17.41**	0.9	X		X
SYNTHETIC-DIR+PP	16.81	0.8	X	X	X

Table 2: Baseline BLEU scores given different requirements: parallel, monolingual, or pivot data. Out-of-vocabulary rates are presented as percentages for each model.

misanalyzed or overly Egyptian*ized*. To achieve this, we extend a DA morphological database with suffix and prefix segmentations, adding a wrapper on top of MADAMIRA to generate the proper segmentation for each analysis. The database extension is automatic and the segmentation is deterministic, following D3 segmentation rules. This allows us to (i) apply this extension to other databases in other dialects that follow the structure of the MADAMIRA database, and (ii) expand our application to dialects that do not have any available analyzers yet.

4 Baseline Models

We use the phrase-based statistical MT platform, Moses (Koehn et al., 2007) to build multiple Egyptian-to-Levantine MT systems: one that only trains on parallel data, another that fabricates pseudo-parallel training data from additional monolingual data, and a third model utilizing pivot data through English. While neural MT has been successfully applied to MSA (Almahairi et al., 2016), we opt for statistical MT as data sparsity and other factors render neural techniques impossible for DA (Zhang et al., 2016). Luong and Manning (2015)'s English-to-Vietnamese neural MT system, for instance, leverages 10 times more parallel data then we use in our experiments, yet still fails to outperform a statistical baseline. Furthermore, their training and testing data is from a single domain with standardized spelling, i.e., limited token:type ratio, which Farajian et al. (2017) suggest should greatly facilitate neural MT performance. Given our sparsity of DA data and lack of spelling conventions, we can neither rely on homogeneous training/testing domains nor low token:type ratios and must resort to statistical MT.

We evaluate the output of our MT systems via BLEU scores (Papineni et al., 2002), comparing them to a single reference in detokenized space. NO-TRANSLATION, scoring 6.48, compares the original, unchanged Egyptian input to the Levantine reference. The results as well as data requirements are reported in Table 2.

4.1 The Direct Model

The most basic statistical system, the DIRECT model can be extended to any dialect pair with parallel data. It is trained only on our BTEC parallel corpus, with some additional monolingual data for language modeling. This model leverages a 2.4 million token 5-gram language model trained using KenLM (Heafield, 2011), consisting of Levantine data from the AOC corpus, BOLT, and BTEC.

Following El Kholy and Habash (2015), we perform word alignment using the grow-diag-final algorithm (Och and Ney, 2003) and we restrict the maximum length of extracted phrases to 8 tokens. Our D3 tokenization is slightly more aggressive than El Kholy and Habash (2015) who use ATB, so we experimented with marginal increases in the maximum allowable phrase length but found them to have no significant effects on performance.

As shown in Table 2, this basic model greatly outperforms the NO-TRANSLATION baseline at 15.44 BLEU, but suffers from a high rate of out-of-vocabulary (OOV) words given that it is only trained on a small amount of parallel data. Furthermore, the model seems to learn noisy weights for many of the phrase pairs it extracts due to the infrequency with which they are encountered during training.

4.2 The Synthetic Model

Inspired by Schwenk and Senellart (2009), we use additional monolingual data to build a SYNTHETIC MT system. First, we use the DIRECT model to translate all of the BOLT Egyptian data to Levantine. Then we build an inverse model identical to the DIRECT model, but from Levantine to Egyptian, and use it to translate the BOLT Levantine data into Egyptian. Finally, we learn a new phrase table from our newly generated parallel corpus consisting of the original 8,000 training sentences, 410,000 BOLT Egyptian-to-generated-Levantine sentences, and 180,000 BOLT Levantine-to-generated-Egyptian sentences.

While Schwenk and Senellart (2009) implement this technique in a slightly different manner for the purpose of domain adaptation, we use it to reduce noise in the phrase table. Due to sparsity of parallel data, the DIRECT model is hard pressed to distinguish good low frequency phrase pairs from bad ones. Adding synthetic data to the model enables it to learn better alignments for low frequency phrase pairs by getting exposure to a variety of different contexts in which such phrases can occur. This system significantly improves over the DIRECT model, scoring 16.75 BLEU, representing our best solution for DA-to-DA MT that does not require pivot data.

4.3 The Phrase Pivot Model

Following El Kholy et al. (2013), we use the BOLT data to phrase pivot through English. Phrase pivoting drastically increases vocabulary coverage; however, it also produces a phrase table with many poorly connected phrase pairs as well as phrase pairs which erroneously translate morpho-syntactic features that cannot be conveyed through morphologically-poor English. The PHRASE PIVOT model addresses the poor connectivity issue by adding El Kholy et al. (2013)'s connectivity strength constraints. These identify how many Egyptian and Levantine tokens in a given Egyptian-to-Levantine-via-English phrase pair can be aligned to each other via corresponding alignments to the same English token.

For example, the noisy Egyptian-to-Levantine phrase pair in Figure 1, would receive a connectivity score of 0.75 from the Egyptian side because 3 of the 4 alignments—those to 'wants', 'to', and 'go'—connect through the English pivot phrase to a Levantine token on the other side. The connectivity score from the Levantine side would be 0.6 because 3 of the 5 Levantine alignments connect all the way through. *hlq* does not count towards the 3 connections because while it connects to the English token 'now', no Egyptian token connects to 'now' from the other side. This example also exhibits the issue that will be addressed in Section 5, that morpho-syntactic properties are not accurately conveyed through morphologically deprived English, as *ʕAyz* and *bd* connect through 'want', though *ʕAyz* implies a masculine subject whereas the suffix of *bd*, *hA*, entails that the subject of the Levantine sentence is in fact third-person feminine.

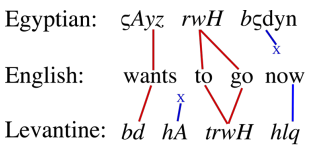

Figure 1: Identifying the connections between an Egyptian phrase and Levantine phrase which were both independently (and noisily) mapped to the same English phrase during pivoting.

This PHRASE PIVOT model can be extended to any DA pair with pivot data, but only marginally outperforms the NO-TRANSLATION baseline at 6.77 BLEU. However, used in conjunction with the DIRECT model, the Direct + Phrase Pivot (DIR+PP) model increases OOV coverage, boosting performance to 17.41 BLEU, almost 2 full BLEU points over the DIRECT baseline. We also re-ran the SYNTHETIC model using DIR+PP to fabricate parallel data instead of DIRECT, however, this did not improve performance. It is possible that the types of DIRECT model errors which are corrected by DIR+PP versus those corrected by SYNTHETIC are similar. Thus, when training on fabricated parallel data, the SYNTHETIC-DIR+PP model may reinforce its own errors more than learn to fix them.

5 Leveraging Morphology and Syntax

The best baseline system, DIR+PP still fails to adequately handle Arabic's rich morphology and syntax, as illustrated by Figure 2, where part-of-speech (POS) is not preserved in the output. A minimally different correct version of the example in Figure 2 would simply replace verbal third-person singular بيفتح *byftH* 'he opens', with the nominal form فتح في *fy ftH* 'in opening'.

Source:				انا عندي مشكلة [[في فتح]] الباب	
	AnA	*ʕndy*	*mšklħ*	[[*fy ftH*]]	*AlbAb*
	I	to-me	problem	[[in opening.N]]	the-door
Output:				انا عندي مشكلة [[بيفتح]] الباب	
	AnA	*ʕndy*	*mšklħ*	[[*byftH*]]	*AlbAb*
	I	to-me	problem	[[with-opens.3MS]]	the-door
Reference:	I'm having trouble opening the door				

Figure 2: Example DIR+PP error where the output does not preserve the POS of the source.

Because Arabic verbs convey person in much finer granularity than do English verbs, which only inflect for third-person singular forms in present tense, many Arabic verb inflections in the source-to-pivot and pivot-to-target phrase tables will be aligned to the same morphologically deprived English verb, e.g., 'opening'. Thus, when the phrase tables are combined via shared English phrases, any given inflected Egyptian verb can be mapped to a large number of Levantine inflections, which, mostly, will not share the same morpho-syntactic properties. In this case, because 'opening', like many 'ing'-suffixed forms in English, can be nominal or verbal, it is not just inflectional morphology that is confused but derivational morphology, as the POS is misinterpreted.

5.1 Addressing Resource Limitations

El Kholy and Habash (2015) use AMEANA (El Kholy and Habash, 2011), an automatic error analysis tool, to determine that definiteness, gender, and number are the features that most frequently contribute to such errors in Hebrew-to-MSA MT. In this work, we were not able to use AMEANA, as it relies on accurate morphological analyses that we cannot produce automatically for Levantine. Furthermore, even if we knew what were the most problematic features for translating Egyptian to Levantine, we might not be able to leverage them, as there is non-trivial noise and Egyptian bias in how the analyses were generated (Section 3).

Our approach focuses instead on identifying features that: (i) MADAMIRA-EGY can recognize relatively accurately (ii) tend to be consistently translated from Egyptian to Levantine. For instance, second-person and third-person verbal forms are frequently orthographically ambiguous in Arabic, making person challenging for our analyzer to correctly identify. Thus,

adding a constraint to the model promoting consistent translation of the person feature value would be useless because we are not likely to know the correct property of person in the first place. Furthermore, if the possible values of a given feature can be translated freely, modeling that feature will be similarly useless. This is often the case with tense, as بشوفك *bšwfk* 'see you' in Egyptian could conceivably be translated as رح شوفك *rH šwfk* in Levantine, changing the value from progressive (realized by the cliticized particle ب *b*) to future tense (realized by the particle رح *rH*).

Without gold, morphologically annotated data in Levantine, we cannot independently measure either our ability to identify morpho-syntactic feature values correctly, or the consistency with which they should be translated. However, we can approximate both jointly. Assuming that Egyptian feature values should correspond to the same feature values on the Levantine side, we measure, for each morpho-syntactic feature, how frequently the realized property on the Egyptian side is aligned to the same property on the Levantine side.

As shown in Table 3, definiteness, gender, number, and POS are the only features which map consistently across aligned tokens in more than 50% of their occurrences throughout the BTEC training set. This suggests that they both can be recognized accurately and are consistently preserved in human translation. Even so, the fact that none of these features map consistently over 80% of the time, suggests that modeling such features will be noisy.

Definiteness	**Number**	**Gender**	**POS**	**Aspect**	**Person**
75	75	62	56	32	29

Table 3: Percentage rates over all training set token alignments at which the feature's values were preserved from Egyptian to Levantine.

5.2 Computing Constraint Scores

Similar to El Kholy and Habash (2015), we design morpho-syntactic constraints by calculating probability distributions from Egyptian to Levantine and vice versa. These reflect how likely each morpho-syntactic property set on one side is to be aligned to each morpho-syntactic property set on the other, based on how often such alignments occured in the training data. Properties sets are defined as the conjunction of values for all morpho-syntactic features under consideration, which for us, include the four most "consistent" features as identified in Table 3: definiteness, number, gender—which were used by El Kholy and Habash (2015)—and also POS (thus, masculine-singular-verb and definite-feminine-noun are property sets). Unaligned tokens are considered to be aligned to a null token on the opposite side, and thus are mapped to an empty property set. We use these probability distributions to add two constraint scores to every phrase pair in the phrase pivot table, one calculated from Egyptian to Levantine and the other, Levantine to Egyptian, as defined in Equations 1 and 2.

$$W_s = \frac{1}{A} \sum_{\forall (i,j) \in a} P(MLE(i)|MLE(j)) \tag{1}$$

$$W_t = \frac{1}{B} \sum_{\forall (i,j) \in b} P(MLE(j)|MLE(i)) \tag{2}$$

We calculate W_s by summing over every alignment a from source token i to target token j (if i is unaligned, j is the null token), the probability of i's property set given j's. While El Kholy and Habash (2015) normalize this sum by the quantity of source tokens, we normalize by the total number of alignments A. Otherwise, many-to-one and one-to-many alignments would bias the scores and enable some to exceed one, making them impossible to interpret as probabilities. The property sets of i and j are determined from the set of all possible property sets for that type (defined as the list of all unique analyses it received over every occurrence in BOLT, AOC, and the BTEC training data) via maximum likelihood estimation, MLE, so as to maximize the likelihood of the source property set given the target property set.

This entails that individual tokens' property sets can be analyzed differenlty from the source side than from the target side. Also, sequences of MLE property sets over multiple tokens on a single side can be syntactically infeasible, e.g., containing 5 consecutive verbs. Thus, we experimented with additional constraints, requiring (i) aligned MLE property sets to have been aligned at least once in the training set (ii) syntactic feasibility on source and target sides independently (iii) alignment of the sequence of property sets on the source side to that on the target side to have appeared at least once in the training set. However, none of these experiments boosted performance, suggesting that MLE inconsistency is not a problematic issue.

W_t is calculated equivalently to W_s from the target side. Adding these constraint weights to each phrase pair in the phrase pivot table, we re-tune the DIR+PP system, re-test, and obtain a statistically significant improvement with a score of 18.03 BLEU on the development set.

We then evaluate the DIRECT, DIR+PP, and Direct + Phrase Pivot with Morpho-syntactic Features (DIR+PP+MORPH) systems on the 2,000 sentence blind test set from our BTEC corpus. The results in Table 4 confirm the utility of our added constraints, as each successive model significantly improves over the last, as in the development set.

Model	Dev	Dev OOV	Test	Test OOV
NO-TRANSLATION	6.48	N/A	6.45	N/A
DIRECT	15.44	4.6	14.61	5.4
DIR+PP	17.41	0.9	16.69	1.0
DIR+PP+MORPH	**18.03**	0.9	**17.48**	1.0

Table 4: Comparing BLEU scores of systems with and without morpho-syntactic features on development and blind test sets. Out-of-vocabulary rates are reported as percentages.

6 Error Analysis

We analyzed the output of the DIR+PP and DIR+PP+MORPH models on 100 development set sentences to investigate the effects of morpho-syntactic features and to identify issues for future work. Table 5 reveals a stark contrast—about a 30% gap in both models—between the quantity of output tokens matching a reference token letter-for-letter (row 3), and the quantity of output tokens manually judged to be acceptable (row 2). Approximately 10% of that 30% gap is due to lack of spelling standards in DA (row 4). Habash et al. (2012a) developed a Conventional Orthography for Dialectal Arabic (CODA) to standardize spelling while preprocessing DA and a prototype system can CODA*fy* Egyptian (Eskander et al., 2013), though no such system is yet available for Levantine. Once developed, we expect such a system to improve MT quality not only by imposing consistent output, but also by reducing sparsity in all translation and language models during training.

The remaining 90% of the 30% gap between exactly matched tokens and tokens judged

to be correct can be approximately split into thirds. One third cannot be directly linked to any reference token (row 7), e.g., tokens in paraphrasal or idiomatic constructions. Another third is tokens which can be linked to a reference token, but take a different root (row 5). The final third are inflectional or derivational variants of the corresponding reference token (row 6). The fact that so many inflectional/derivational variants are judged correct demonstrates that morpho-syntactic modeling is necessarily noisy as property sets are frequently not preserved, even in acceptable translations. On the other hand, the success of DIR+PP+MORPH suggests that some features' properties tend to be preserved through translation or at least altered predictably, as otherwise, the system would not benefit from modeling them.

		DIR+PP	DIR+PP MORPH
1	**Words**	665	670
2	**Words Judged Correct**	**569 (85.6)**	**594 (88.7)**
3	Exact Match	377 (56.7)	382 (57.0)
4	CODA Variant	23 (3.5)	25 (3.7)
5	Different Root	47 (7.1)	51 (7.6)
6	Different Properties	61 (9.2)	65 (9.8)
7	Otherwise Different	61 (9.2)	71 (10.7)
8	**Words Judged Incorrect**	**96 (14.4)**	**76 (11.3)**
9	Morpho-syntactic Properties	49 (7.4)	41 (6.1)
10	Other Problems	47 (7.1)	35 (5.2)
11	**Properties**: Modeled	33 (5.0)	26 (3.9)
12	Not Modeled	16 (2.4)	15 (2.2)
13	**Other**: Wrong Word Sense	18 (2.7)	17 (2.5)
14	Apparent Phrasal Issue	13 (2.0)	7 (1.0)
15	Unclear Reason	13 (2.0)	7 (1.0)
16	OOV	2 (0.3)	3 (0.4)
17	Copies Egyptian	1 (0.2)	1 (0.1)
18	**Word Error Reduction**	**N/A**	**(20.2)**
19	**Sentences**	100	100
20	Correct Sentences	48	55
21	Sentence Error Reduction	N/A	(13.5)

Table 5: Comprehensive manual error analysis of 100 sentences from the development set. Values within parentheses are percentages.

6.1 Direct Effects of Added Features

The second major insight of the error analysis is that the error reduction from adding morpho-syntactic constraints is far more significant (20.2%, row 18), than the improvement registered by the automatic BLEU scores. Example sentences illustrating some of these improvements are contained in Figure 3. For 7.4% of the tokens DIR+PP outputs, its only mistake is mis-representing one or more morpho-syntactic property (row 9). Comparing that to 6.1% for DIR+PP+MORPH (row 9), the new model makes a 17% error reduction in the area it is designed to improve. Furthermore, essentially all of this improvement takes place in sentences where DIR+PP+MORPH corrects a mistake involving a feature we model: definiteness, gender, number, or POS. For example, the definite article is correctly added to the word الحقيقة *AlHqyqħ* 'the truth' in Figure 3a.

	ENGLISH:	Actually, inside it [...] Do you mind if I take it out?							
	REFERENCE:	بالحقيقة في بالبو [...] عندك مانع ازا شلتو؟							
		bAlHqyqħ	*fy*	*bAlbw*	[...]	*ʕndk*	*mAnʕ*	*AzA*	*šltw?*
		with-the-truth	exists	in-heart-its	[...]	to-you	problem	if	take-1SPast-it
(a)	**DIR+PP:**	حقيقة* في بالبا [...] عندك مانع اذا طلعو ؟*							
		*Hqyqħ**	*fy*	*bAlbA*	[...]	*ʕndk*	*mAnʕ*	*AðA*	*Tlʕw?**
		truth*	exists	in-heart-its	[...]	to-you	problem	if	remove.3SPast-it*
	DIR+PP+MORPH:	الحقيقة ، هوي بقلبو [...] عندك مانع اذا شلتو؟							
		AlHqyqħ	*hwy*	*bqlbw*	[...]	*ʕndk*	*mAnʕ*	*AðA*	*šltw?*
		the-truth	it	in-heart-its	[...]	to-you	problem	if	take.1SPast-it?

	ENGLISH:	I'll bring one right away				
	REFERENCE:	رح جيب واحد هلا				
		rH	*jyb*	*wAHd*	*hlA*	
		will	bring.1S	one	now	
(b)	**DIR+PP:**	بالجيبة* واحد هلق				
		*bAljybħ**	*wAHd*	*hlq*		
		with-the-pocket*	one	now		
	DIR+PP+MORPH:	رح جيب واحد هلق				
		rH	*jyb*	*wAHd*	*hlq*	
		will	bring.1S	one	now	

	ENGLISH:	What kind of fruit do you have?					
	REFERENCE:	اي نوع فواكي عندك؟					
		Ay	*nwʕ*	*fwAky*	*ʕndk?*		
		which	kind	fruit	at-you		
(c)	**DIR+PP:**	عندك اي نوع من الفاكهة شو؟*					
		ʕndk	*Ay*	*nwʕ*	*mn*	*AlfAkhħ*	*šw?**
		at-you	which	kind	from	the-fruit	what*
	DIR+PP+MORPH:	عندك اي نوع من الفاكهة؟					
		ʕndk	*Ay*	*nwʕ*	*mn*	*AlfAkhħ?*	
		at-you	which	kind	from	the-fruit	

Figure 3: Example translation errors (marked with *) corrected by adding morpho-syntactic constraints to the model.

6.2 Indirect Effects of Added Features

Surprisingly, most of the overall error reduction actually comes from mistakes other than misrepresentation of morpho-syntactic properties, as such mistakes decrease by 26%, from 7.1% to 5.2% (row 10). The morpho-syntactic constraints seem to teach the model about syntax at the phrase level, as sentences like Figure 3b are corrected, which were originally over-chunked into small phrases by DIR+PP. بالجيبة *bAljybħ* 'with the pocket' is likely a misanalysis of the source word هاجيب *hAjyb* 'I'll get' translated as a single-word phrase, as it would have made for an infrequent or non-existent bigram or trigram when combined with the following words in the output. The DIR+PP model likely had access to longer phrase pairs such as هاجيب واحد *hAjyb wAHd* 'I'll get one' mapping to رح جيب واحد *rH jyb wAHd*—the corresponding reference phrase—but likely did not select it because longer phrases are inherently less frequent,

i.e., noisier to model.

Morpho-syntactic constraints enable DIR+PP+MORPH to increasingly select longer, infrequent phrase pairs by distinguishing those that are morpho-syntactically feasible from competing shorter alternatives. Translating larger phrasal chunks leads to more fluent output by reducing opportunities to incorrectly chunk phrase boundaries. This is why much of the error reduction does not appear to be, superficially at least, related to morpho-syntactic features targeted by DIR+PP+MORPH.

Figure 3c exhibits another benefit of morpho-syntactic features, as DIR+PP+MORPH often corrects the insertion of gratuitous words, here, شو *šw* 'what'. Such features teach the model that certain POS's are less likely to align to the null token, even if the language model favors the sequence with the gratuitous token monolingually.

7 Conclusion and Future Work

In this work, we presented the second ever evaluated Arabic DA-to-DA MT effort. The subject has not previously received serious attention due to lack of naturally occurring parallel data, though DA is widely spoken and dialects are frequently mutually unintelligible, exhibiting comparable linguistic variation to the Romance languages. Our results suggest that modeling morphology and syntax can significantly improve DA-to-DA MT despite data sparsity. However, optimizing models under such circumstances requires careful consideration of the linguistic differences between dialects and careful tailoring and implementation of all available data and resources.

Given that many DA pairs may not have pivot data available, the most pressing future work is to develop a dialect-agnostic tokenizer and analyzer which does not suffer from the Egyptian bias that ours does. This will reduce data sparsity regardless of the nature of the low-resourced MT settings for any DA pair, and it will enable better morpho-syntactic modeling.

Additionally, improving on the dialect identification work of Diab et al. (2010), Zaidan and Callison-Burch (2011), and Elfardy and Diab (2013) will enable us to collect more monolingual data. This data is not only useful for language modeling but can also be mined for comparable sentences to augment the parallel training set. The process typically involves using metadata (Resnik and Smith, 2003) and seed data (Munteanu and Marcu, 2005) to identify pairs of related sentences or phrases in the source and target languages (Cettolo et al., 2010; Max et al., 2012). These are then iteratively classified via expectation maximization with phrases identified as parallel being added to the seed data (Dong et al., 2015). Models trained thusly produce noisy phrase pairs, often imperfectly modeling morpho-syntactic property sets. Thus, the same morpho-syntactic constraints that improved DIR+PP+MORPH can be adapted to improve MT via comparable corpora.

8 Acknowledgments

This publication was made possible by grant NPRP 7-290-1-047 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors. The first author was funded by the Boren Fellowship program.

References

Al-Shargi, F., Kaplan, A., Eskander, R., Habash, N., and Rambow, O. (2016). Morphologically annotated corpora and morphological analyzers for Moroccan and Sanaani Yemeni Arabic. In *10th Language Resources and Evaluation Conference (LREC 2016)*.

Almahairi, A., Cho, K., Habash, N., and Courville, A. (2016). First result on Arabic neural machine translation. *arXiv preprint arXiv:1606.02680*.

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese–Spanish machine translation. In *International Workshop on Computational Processing of the Portuguese Language*, pages 50–59.

Bouamor, H., Habash, N., and Oflazer, K. (2014). A multidialectal parallel corpus of Arabic. In *LREC*, pages 1240–1245.

Cettolo, M., Federico, M., and Bertoldi, N. (2010). Mining parallel fragments from comparable texts. In *IWSLT*, pages 227–234.

Chiang, D., Diab, M., Habash, N., Rambow, O., and Shareef, S. (2006). Parsing Arabic Dialects. In *Proceedings of EACL*, Trento, Italy. EACL.

Corbí Bellot, A. M., Forcada, M. L., Ortiz Rojas, S., Pérez-Ortiz, J. A., Ramírez Sánchez, G., Sánchez-Martínez, F., Alegría Loinaz, I., Mayor Martínez, A., Sarasola Gabiola, K., et al. (2005). An open-source shallow-transfer machine translation engine for the Romance languages of Spain. *Proceedings of the 10th Conference of the European Association for Machine Translation*, pages 79–86.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. In *LREC*, pages 241–245.

Diab, M., Habash, N., Rambow, O., Altantawy, M., and Benajiba, Y. (2010). Colaba: Arabic dialect annotation and processing. In *LREC Workshop on Semitic Language Processing*, pages 66–74.

Diab, Mona T and Al-Badrashiny, Mohamed and Aminian, Maryam and Attia, Mohammed and Elfardy, Heba and Habash, Nizar and Hawwari, Abdelati and Salloum, Wael and Dasigi, Pradeep and Eskander, Ramy. (2014). Tharwa: A Large Scale Dialectal Arabic-Standard Arabic-English Lexicon. In *LREC*, pages 3782–3789.

Dong, M., Liu, Y., Luan, H.-B., Sun, M., Izuha, T., and Zhang, D. (2015). Iterative learning of parallel lexicons and phrases from non-parallel corpora. In *IJCAI*, pages 1250–1256.

Durrani, N., Al-Onaizan, Y., and Ittycheriah, A. (2014). Improving Egyptian-to-English SMT by mapping Egyptian into MSA. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 271–282.

El Kholy, A. and Habash, N. (2011). Automatic error analysis for morphologically rich languages. In *MT Summit XIII*.

El Kholy, A. and Habash, N. (2012). Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*, 26(1-2):25–45.

El Kholy, A. and Habash, N. (2015). Morphological Constraints for Phrase Pivot Statistical Machine Translation. In *Machine Translation Summit*, Miami.

El Kholy, A., Habash, N., Leusch, G., Matusov, E., and Sawaf, H. (2013). Language independent connectivity strength features for phrase pivot statistical machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 412–418.

Elfardy, H. and Diab, M. (2013). Sentence Level Dialect Identification in Arabic. In *Proceedings of the Association for Computational Linguistics*, pages 456–461, Sofia, Bulgaria.

Eskander, R., Habash, N., Rambow, O., and Pasha, A. (2016). Creating resources for Dialectal Arabic from a single annotation: A case study on Egyptian and Levantine. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3455–3465, Osaka, Japan.

Eskander, R., Habash, N., Rambow, O., and Tomeh, N. (2013). Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Farajian, M. A., Turchi, M., Negri, M., Bertoldi, N., and Federico, M. (2017). Neural vs. phrase-based machine translation in a multi-domain scenario. *EACL 2017*, page 280.

Habash, N., Diab, M., and Rambow, O. (2012a). Conventional Orthography for Dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 711–718, Istanbul, Turkey.

Habash, N., Eskander, R., and Hawwari, A. (2012b). 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.

Habash, N. and Hu, J. (2009). Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 173–181, Athens, Greece.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On Arabic Transliteration. In van den Bosch, A. and Soudi, A., editors, *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Habash, N., Zalmout, N., Taji, D., Hoang, H., and Alzate, M. (2017). A parallel corpus for evaluating machine translation between arabic and european languages. *EACL 2017*, page 235.

Habash, N. Y. (2010). *Introduction to Arabic natural language processing*, volume 3. Morgan & Claypool Publishers.

Hajič, J., Smrž, O., Zemánek, P., Šnaidauf, J., and Beška, E. (2004). Prague Arabic Dependency Treebank: Development in data and tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, pages 110–117, Cairo, Egypt.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Jarrar, M., Habash, N., Akra, D., and Zalmout, N. (2014). Building a corpus for Palestinian Arabic: a preliminary study. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 18–27.

Khalifa, S., Habash, N., Abdulrahim, D., and Hassan, S. (2016). A large scale corpus of Gulf Arabic. *arXiv preprint arXiv:1609.02960*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86.

Koehn, P., Birch, A., and Steinberger, R. (2009). 462 Machine Translation Systems for Europe. In *Proceedings of MT Summit*, Ottawa, Canada.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Luong, M.-T. and Manning, C. D. (2015). Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*.

Maamouri, M., Bies, A., Kulick, S., Ciul, M., Habash, N., and Eskander, R. (2014). Developing an Egyptian Arabic treebank: Impact of dialectal morphology on annotation and tool development. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Max, A., Bouamor, H., and Vilnat, A. (2012). Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 721–731.

McNeil, K. and Faiza, M. (2011). Tunisian Arabic corpus: Creating a written corpus of an unwritten language. In *Workshop on Arabic Corpus Linguistics (WACL)*.

Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., and Smaili, K. (2015). Machine translation experiments on PADIC: A parallel Arabic dialect corpus. In *The 29th Pacific Asia conference on language, information and computation*.

Munteanu, D. S. and Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Muraki, K. (1987). PIVOT: Two-phase machine translation system. In *MT Summit Manuscripts and Program*, pages 81–83.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of LREC*, Reykjavik, Iceland.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Sadat, F. and Habash, N. (2006). Combination of Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1–8.

Sajjad, H., Darwish, K., and Belinkov, Y. (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.

Salloum, W. and Habash, N. (2011). Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the first workshop on algorithms and resources for modelling of dialects and language varieties*, pages 10–21.

Salloum, W. and Habash, N. (2012). Elissa: A Dialectal to Standard Arabic Machine Translation System. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012): Demonstration Papers*, pages 385–392, Mumbai, India.

Sawaf, H. (2010). Arabic dialect handling in hybrid machine translation. In *Proceedings of AMTA*, Denver, Colorado.

Schwenk, H. and Senellart, J. (2009). Translation model adaptation for an Arabic/French news translation system by lightly-supervised training. In *MT Summit*.

Tachicart, R. and Bouzoubaa, K. (2014). A hybrid approach to translate Moroccan Arabic dialect. In *Intelligent systems: Theories and applications (sita-14), 2014 9th international conference on*, pages 1–5. IEEE.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *LREC*, pages 147–152.

Tyers, F. M., Font, H. A. i., Fronteddu, G., and Martín-Mor, A. (2017). Rule-based machine translation for the Italian-Sardinian language pair. In *The Prague Bulletin of Mathematical Linguistics No. 108*, pages 221–232.

Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *HLT-NAACL*, pages 484–491.

Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic.

Zaidan, O. F. and Callison-Burch, C. (2011). The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the Association for Computational Linguistics*, Portland, Oregon, USA.

Zbib, R., Malchiodi, E., Devlin, J., Stallard, D., Matsoukas, S., Schwartz, R., Makhoul, J., Zaidan, O. F., and Callison-Burch, C. (2012). Machine translation of Arabic dialects. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 49–59.

Zhang, D., Kim, J., Crego, J. M., and Senellart, J. (2016). Boosting neural machine translation. *CoRR*, abs/1612.06138.