Translate, Predict or Generate: Modeling Rich Morphology in Statistical Machine Translation

Ahmed El Kholy and **Nizar Habash**
Center for Computational Learning Systems, Columbia University
475 Riverside Drive New York, NY 10115
{akholy,habash}@ccls.columbia.edu

Abstract

We compare three methods of modeling morphological features in statistical machine translation (SMT) from English to Arabic, a morphologically rich language. Features can be modeled as part of the core translation process mapping source tokens to target tokens. Alternatively these features can be generated using target monolingual context as part of a separate generation (or post-translation inflection) step. Finally, the features can be predicted using both source and target information in a separate step from translation and generation. We focus on three morphological features that we demonstrate through a manual error analysis to be most problematic for English-Arabic SMT: gender, number and the determiner clitic. Our results show significant improvements over a state-of-the-art baseline (phrase-based SMT) of almost 1% absolute BLEU on a medium size training set. Our best configuration models the determiner as part of core translation and predicts gender and number separately, and handles the rest of the features through generation.

1 Introduction

Translation into English has been the focus of many research efforts in Statistical Machine Translation (SMT). However, recently, translation into other languages has been receiving increasing attention, especially translation into morphologically rich languages (Sarikaya and Deng, 2007; Elming and Habash, 2009; Yeniterzi and Oflazer, 2010).

One of the main issues in SMT is the sparsity of parallel data for many language pairs especially when the source or target language is morphologically rich. Morphological richness comes with many challenges and the severity of these challenges increases when translating from a morphologically poor language to a morphologically richer language.

In this paper, we address these challenges through different modeling methods.[1] In our approach, morphological features can be modeled as part of the core translation process mapping source tokens to target tokens. Alternatively these features can be generated using target monolingual context as part of a separate generation (or post-translation inflection) step. Finally, the features can be predicted using both source and target information in a separate step before generation. We focus in our experiments on English-Arabic SMT and we work on three morphological features that we found, through a manual error analysis, to be most problematic for English-Arabic SMT: gender, number and the determiner clitic. Our results show improvements over a state-of-the-art baseline (phrase-based SMT) of almost 1% absolute BLEU on a medium size training set of 4M words. Our best configuration models the determiner as part of core translation, predicts gender and number features separately, and handles the rest of the features through generation. We test our approach on a blind test set and we got the same relative improvements across the different systems. However, when scaling up the data set, the advantage of using morphological modeling disappears, which is not surprising.

2 Related Work

There have been numerous efforts studying the effect of applying morphological processing or using morphological information on SMT quality. In one approach, Factored SMT, morphological features can be modeled jointly as factors in the trans-

[1]This work was funded by a Google research award.

lation process (Koehn et al., 2007). These factors can be used in different translation and generation expansion steps. One of the main drawbacks of this approach is the combinatorial expansion of the number of translation options.

Another approach is to model translation and morphology independently in a sequential manner. A common method within this approach is to morphologically preprocess the training data before training the translation models, e.g., morphological tokenization of clitics (Habash and Sadat, 2006; Oflazer and Durgar El-Kahlout, 2007; Badr et al., 2008). Tokenization reduces sparsity of the data and increases the symmetry between source and target, which in return improves the quality of the translation. There is a large space of different tokenization schemes for Arabic. In our experiments, we use the Penn Arabic Treebank (PATB) tokenization scheme which was shown in previous effort by El Kholy and Habash (2010a) to perform well when translating into Arabic. As a result of tokenization, a post-processing step is needed to recombine (detokenize) the clitics back to the word. This is a somewhat complex task involving several orthographic and morphological adjustments (El Kholy and Habash, 2010b).

Another method related to our approach is using an independent morphological prediction component such as used by Minkov et al. (2007) and Toutanova et al. (2008). They use maximum entropy models for inflection prediction. Unlike our approach, they predict inflected word forms directly without going into a fine grained morphological feature prediction as we do. One of the main drawbacks of their approach is that they use stems as their base for translation instead of lemmas (see Section 3.1). On average, a lemma in Arabic could have two stems so using lemmas can make the data less sparse and make the translation model tighter. There is also work by Clifton and Sarkar (2011) where they do segmentation and morpheme prediction. They also use stems as their basic word form.

3 English-Arabic SMT Challenges

In this section, we discuss the challenges of the English-Arabic language pair in the context of MT. We also provide two error analyses that helped define the scope of our work and motivated our experimental setup.

3.1 Linguistic Facts

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

One aspect of Arabic's complexity has to do with its orthography which often omits short-vowel diacritics. As a result, ambiguity is rampant. Another aspect of Arabic that contributes to this complexity is its various attachable clitics which include conjunction proclitics, e.g., +و *w*+ 'and', particle proclitics, e.g., +ل *l*+ 'to/for', the definite article +ال *Al*+ 'the', and the class of pronominal enclitics, e.g., +هم *hm*+ 'their/them'. Beyond these clitics, Arabic words inflect for person (PER), gender (GEN), number (NUM), aspect (ASP), mood (MOD), voice (VOX), state (STT) and case (CAS). This morphological richness leads to thousands of inflected forms per lemma and a high degree of ambiguity: about 12 analyses per word, typically corresponding to two lemmas on average (Habash, 2010). The PATB tokenization scheme (Maamouri et al., 2004) which we use in our baseline and all experiments separates all clitics except for the determiner clitic *Al*+ (DET).

Arabic also has complex morpho-syntactic agreement rules in terms of GEN and NUM within specific constructions such as nouns with their adjectives and verbs with their subjects (Alkuhlani and Habash, 2011). The DET in Arabic is used to distinguish different syntactic constructions such as the possessive or adjectival modification.

English on the other hand barely inflects for NUM and tense and for PER in a limited context. The NUM feature in Arabic has more values (dual) than English. GEN in English is not expressed morphologically. When translating from English into Arabic, we expect to be able to model shared morphological features more than absent features or features expressed only syntactically in English or Arabic, e.g., the possessive construction.

3.2 Automatic Error Analysis

We conducted an error analysis of our baseline system on our development set (MT05) using an open-source tool for error analysis of natural lan-

guage processing tasks targeting morphologically rich languages (El Kholy and Habash, 2011). The tool aligns words in the output and the reference if they share the same lemma. Each output word receives a matching category based on the reference word it is paired with. If the output and reference words have same form, the category is Exact Match, otherwise, it is Lemma Match. Unpaired output words are tagged Unmatchable. The tool also produces detailed statistics on morphological errors. Exact Match cases are 59.0% and Lemma Match cases are 13.3%. Among Lemma Match cases, DET is the biggest single feature error. The PATB clitics errors (53.6%) together with DET (29.7%), GEN (12.8%) and NUM (10.8%) are the biggest culprits overall. This analysis suggests targeting them may be most beneficial.

3.3 Manual Error Analysis

We also performed a manual error analysis on a hundred sentences from the output of the MT05 set translated with the baseline system. Exact Match cases are 57% and Lemma Match cases are 15%. Among Unmatchable cases, 21.4% of the words have good paraphrases. We looked at the morphological errors that affect adequacy and fluency. We define a morphological adequacy error as the mistranslation of a certain morphological feature conveying a different meaning from the English. Morphological fluency errors are morpho-syntactic disagreements in the Arabic output. Table 1 summarizes our findings. In terms of adequacy, DET along with NUM are the biggest culprits overall. In terms of fluency, GEN is far worse than any other feature which highlights its importance. Another important observation is that the union of the words which affect both fluency and/or adequacy are almost 6.5% which defines the upper limit of words that can improve through morphological modeling.

4 Approach

In our approach, the process of translating English words to Arabic words is broken into a pipeline consisting of four steps:

- **Lexical Translation** from English words to tokenized Arabic lemmas and any subset of Arabic linguistic features.
- **Morphology Prediction** of linguistic features to inflect Arabic lemmas.
- **Morphology Generation** of inflected Arabic tokens from Arabic lemmas and any subset of Arabic linguistic features.
- **Detokenization** of inflected Arabic tokens into surface Arabic words.

Arabic tokenization and lemmatization are done before training the translation models. Both lexical translation and generation are implemented as phrase-based SMT systems (Koehn et al., 2007). Morphology prediction is an optional step implemented using a supervised discriminative learning model. Generation can be done from lemmas and any subset of Arabic inflectional features. Detokenization simply stitches the words and clitics together as a post-processing step (Badr et al., 2008; El Kholy and Habash, 2010a).

We follow numerous previously published efforts on the value of tokenization for English-Arabic SMT (Badr et al., 2008; El Kholy and Habash, 2010a; Al-Haj and Lavie, 2010) and focus on the question of how to improve the translation of tokenized words using deeper representations, namely lemmas and features. Within our framework, we can model the translation of different Arabic linguistic features as part of the lexical translation step, as part of the generation step, or model them using an independent morphology prediction step. Some features, such as clitics, can be modeled well through simple tokenization and detokenization (which can be thought of as part of lexical translation).

We build on a previous effort in improving the quality of the English-to-Arabic translation through Arabic tokenization (El Kholy and Habash, 2010a). We use the best performing tokenization scheme (PATB) and the best detokenization technique on the output as our baseline. Consequently, in this paper we focus on the first three components of the pipeline and we keep the tokenization a constant across all experiments. We study different options of including three morphological features (GEN, NUM and DET) in the first three steps of the pipeline and their implications on the quality of English-to-Arabic SMT. We discuss the three steps in the following subsections.

4.1 Lexical Translation

Lexical translation is the first step in our decoding pipeline. It is trained on pre-processed text: tokenized, lemmatized and disambiguated Arabic words and English words (with limited processing) and their POS tags. We use an SMT system to translate from English words (ENGWORD) and POS tags (POS) to tokenized Arabic lemmas (AR-

Words with Morphological Errors Affecting		Percentage of Morphology Error Type								
		Tokenized PATB Clitics			Non-tokenized Morphological Features					
		CONJ	PART	PRON	DET	PER	ASP	GEN	NUM	CAS
Adequacy	2.6	3.6	7.3	7.3	38.2	1.8	7.3	12.7	30.9	0.0
Fluency	5.1	10.8	13.5	7.2	18.9	0.9	0.9	41.4	19.8	2.7
Adequacy ∪ Fluency	6.5	9.6	13.0	7.6	**26.8**	1.4	3.4	**34.2**	**17.8**	2.0

Table 1: Column two presents the percentage of words with morphological errors that affect the adequacy and fluency of the translation quality. Starting from column three till the end are percentages of the error contributed by each morphological features. Since multiple errors can occur, these values overlap.

ALEM) plus zero or more morphological features. We use an abstract representation for the morphological features so that each word is represented as a lemma and a set of feature-value pairs. Table 2 shows a sample sentence in the above-mentioned representations. This way we simplify the translation task by targeting a less complex output. The key point here is to keep the morphological features that help the translation task and then try to generate the rest of the morphological features and inflected forms in later steps. The output of lexical translation is input to the morphological generation step directly or is first enriched by additional morphological features predicted in the morphology prediction step.

4.2 Morphology Prediction

Morphology prediction takes the output of lexical translation and tries to enrich it by predicting one or more morphological features. Unlike Toutanova et al. (2008), who predict full inflected forms and Clifton and Sarkar (2011) who predict morphemes, we predict morphological feature. This task is, in sense, a form of POS tagging. However, unlike typical tagging, which is done on fully inflected word forms, this task is applied to uninflected or semi-inflected forms – lemmas with zero or more morphology features. As such, we do not expect it to do as well as normal POS tagging/morphology disambiguation for Arabic (Habash and Rambow, 2005).

We use a Conditional Random Field (CRF) toolkit (Lafferty et al., 2001) to train a prediction module with a variety of learning features (not to be confused with the tagged linguistic features). We also make use of the alignment information produced by the MT system in the lexical translation step to get the equivalent aligned English word of each translated word. We then use this information in addition to some syntactic information on the English side as CRF learning features.

We group the CRF learning features into two sets: *Basic* and *Syntax*. The *Basic* features consist of the Arabic output from the lexical translation step (lemma plus certain features), the equivalent aligned English word, English POS and English context (+/- two words). The *Syntax* features consist of the English parent word in a dependency tree, the dependency relation and the equivalent Arabic output word of the English parent. English is parsed using the Stanford Parser (Klein and Manning, 2003).

In training the CRF model, we use the same data used in training the lexical translation step (Section 5). We create three datasets from this data. The first is the original gold data where we train the CRF module on clean Arabic text and gold feature values that are determined using a state-of-the-art POS tagger for Arabic (Habash and Rambow, 2005). Although the automatic tagging does produce errors, we still call this data set *gold* since the Arabic is correctly inflected naturally occurring text. The second dataset is created by translating the whole data using the translation model created by the lexical translation step. The intuition here is to model lexical translation errors by training the CRF models on data similar in quality to its expected input. The last dataset is the combination of gold and translated dataset.

Table 3 shows the accuracy of the CRF module on a test set of 1000 sentences. CRF in general achieves a high accuracy across the different training datasets and the different training parameters. Using translated data does not outperform using gold data; however, the accuracy of predicting NUM and GEN seems to benefit from adding the translated data to the gold data. That could be explained by the fact that NUM and GEN are more affected by translation adequacy unlike DET which is more coupled with translation fluency. Overall the results are about 10-14% absolute lower than MADA (Habash and Rambow, 2005) tagging of the same features on fully inflected text; and are 20-30% absolute better than a degenerate baseline using the most common feature value.

Representation	Example
ENGWORD	saddam hussein 's half-brother refuses to return to iraq
ENGWORD)+POS	saddam#NN hussein#NN 's#POS half-brother#NN refuses#VBZ to#TO return#VB to#TO iraq#NN
ARALEM	Âax γayor šaqiyq li+ Sad∼Am Husayon rafaD ςawodaℏ Iilaý ςirAq
ARALEM+DET	Âax#det γayor#0 šaqiyq#det li+#na Sad∼Am#0 Husayon#0 rafaD#0 ςawodaℏ#det Ăilaý#na ςirAq#det
Arabic Tokenized	AlÂx γyr Alšqyq l+ SdAm Hsyn yrfD Alςwdℏ Ălý AlςrAq
Arabic Script	الأخ غير الشقيق لصدام حسين يرفض العودة الى العراق

Table 2: A sample sentence showing the different representations used in our experiments.

The morphology prediction step produces a lattice with all possible feature values each having an associated confidence score. The morphology generation module discussed next will decide on the best option.

Prediction Training		Predicted Feature Accuracy		
Data Set	**Model**	GEN	NUM	DET
Gold	**Basic**	84.65	88.76	**88.00**
	Basic+Syntax	84.22	89.11	87.85
Translated	**Basic**	84.46	86.11	85.98
	Basic+Syntax	84.08	86.79	85.41
Gold +Translated	**Basic**	**85.96**	89.43	87.40
	Basic+Syntax	85.49	**89.52**	86.91

Table 3: Accuracy (%) of feature prediction starting from Arabic lemmas. A most-common-tag degenerate baseline would yield 67.4%, 70.6% and 59.7% accuracy for GEN, NUM, and DET, respectively. Reported MADA classification accuracy starting from fully inflected Arabic is as follows: GEN 98.2% , NUM 98.8%, DET 98.3% (Habash and Rambow, 2005).

4.3 Morphology Generation

Morphology generation maps Arabic lemmas (ARALEM) plus morphological features to Arabic inflected forms. This step is implemented as an SMT system that translate from a deeper linguistic representation to a surface representation of each token. This step is conceptually similar to the generation expansion component in Factored SMT, but it is implemented as a complete SMT system. The main advantage of this approach is that the training data is not restricted to parallel corpora. We can use all the monolingual data we have in building the system. For more details, see (El Kholy and Habash, 2012).

To evaluate the performance of this approach in generating Arabic inflected forms, we built several SMT systems translating from ARALEMs plus zero or more morphological features to Arabic inflected form. We use the same tools and setup as discussed in Section 5. Table 4 shows the BLEU scores of generating the MT05 set starting from Arabic lemmas plus different morphological features (GEN, NUM, DET), and their combinations. As expected, the more features are included the better the results. Here comes the trade off between the lexical translation quality and morphological generation. The BLEU scores are very high because the input is golden in terms of word order and lemma choice. These scores should be seen as the upper limit on correctness that can be expected from this step, rather than its actual performance in an end-to-end pipeline.

Gold Generation Input	BLEU%
ARALEM	82.19
ARALEM+DET	86.62
ARALEM+NUM	86.89
ARALEM+GEN	87.32
ARALEM+GENNUM	90.18
ARALEM+GENNUMDET	94.77

Table 4: Results of generation from gold ARALEM plus different sets of morphological features. Results are in (% BLEU) on the MT05 set.

The morphology generation step can take the output of lexical translation directly or after predicting certain morphological features using the morphology prediction step.

5 Experiments

In this section, we present our results comparing the modeling of GEN, NUM and DET features, first as part of lexical translation versus morphological generation, and then as part of morphological prediction versus morphological generation. We also present results on a blind test set MT06, a much larger training corpus, and discuss our findings.

5.1 Experimental Setup

All of the training data we use is available from the Linguistic Data Consortium (LDC).[2] We use an English-Arabic parallel corpus of about 142K sentences and 4.4 million words for translation model training data. The parallel text includes Arabic News (LDC2004T17),

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

eTIRR (LDC2004E72), English translation of Arabic Treebank (LDC2005E46), and Ummah (LDC2004T18). Word alignment is done using GIZA++ (Och and Ney, 2003). For language modeling, we use 200M words from the Arabic Gigaword Corpus (LDC2007T40) together with the Arabic side of our training data. We used 5-grams for all LMs implemented using the SRILM toolkit (Stolcke, 2002).

MADA is used to tokenize the Arabic text and produce lemmas and their accompanied morphological features. English preprocessing simply includes down-casing, separating punctuation and splitting off "'s".

All experiments are conducted using the Moses phrase-based SMT system (Koehn et al., 2007). The decoding weight optimization was done using a set of 300 sentences from the 2004 NIST MT evaluation test set (MT04). The tuning is based on tokenized Arabic without detokenization. We use a maximum phrase length of size 8. We report results on the 2005 NIST MT evaluation set (MT05). These test sets were created for Arabic-English MT and have four English references. We arbitrarily picked the first English reference to be source and used the Arabic source as the only reference. We evaluate using BLEU-4 (Papineni et al., 2002).

Our baseline replicates the work of El Kholy and Habash (2010a), who determined that tokenizing Arabic into the PATB tokenization scheme is optimal for phrase-based SMT models. The baseline BLEU score is 29.48% using exactly the same data sets used in the rest of the experiments.

5.2 Translation vs. Generation

We compare the performance of translating English and English plus POS into Arabic lemmas plus different morphological feature combinations followed by generation of the final Arabic inflected form using the morphology generation step directly under the same conditions. The results are presented in Table 5. The best performer across all conditions is translating English words to Arabic lemmas plus DET. This is the only setup that beats the baseline system. The difference in BLEU scores between this setup and the baseline is statistically significant above the 95% level. Statistical significance is computed using paired bootstrap resampling (Koehn, 2004). This shows the importance of DET in lexical translation. English POS oddly does not help. This is perhaps a result of the added sparsity in how we modeled them (as ENGWORD+POS). It is possible a factored MT model can give different results. We plan to explore this question in the future.

Input	A'	BLEU%
ENGWORD	ARALEM	**29.47**
ENGWORD+POS	ARALEM	29.26
ENGWORD	ARALEM+NUM	28.96
ENGWORD+POS	ARALEM+NUM	28.52
ENGWORD	ARALEM+GEN	28.81
ENGWORD+POS	ARALEM+GEN	28.65
ENGWORD	ARALEM+DET	**30.13**
ENGWORD+POS	ARALEM+DET	29.33
ENGWORD	ARALEM+GENNUM	28.82
ENGWORD+POS	ARALEM+GENNUM	28.65
ENGWORD	ARALEM+GENNUMDET	29.19
ENGWORD+POS	ARALEM+GENNUMDET	29.00

Table 5: End-to-end MT results for different settings of English input and Intermediate Arabic. Results are in (% BLEU) on our MT05 set.

5.3 Prediction vs. Generation

We compare results of two translation settings and a variety of added predicted features. The results are presented in Table 6. We can see from the results that using predicted GEN by itself does not help across the board yet it could be helpful when combined with other features. It also seems that predicting NUM when lexical translation is done with lemmas only helps the performance but that is not the case when the lexical translation is done using Lemma plus DET. Another observation is that combining GEN and NUM degrades the overall performance more than the GEN by itself; however, we get the best scores when DET is combined with them. This shows that some synergies come out when different features are combined together even if they perform badly on their own. The only fact that seems very robust is that translating English to Lemma plus DET and then predicting both GEN and NUM gives the highest scores. Predicting features using models trained on translated texts seem to also consistently do better than using models that are trained on original Arabic. The best result obtained is statistically significant compared with the best reported score in the previous section (ARALEM+DET translation).

5.4 Blind Test

We performed a blind test using the 2006 NIST MT evaluation set (MT06) and compared the results to (MT05). MT06 is a harder set to translate than MT05. However, the relative performance is

Translation	ENGWORD→ARALEM					ENGWORD→ARALEM+DET		
No Prediction	29.47					30.13		
Prediction Training	**Predicted Morphological Features**							
	GEN	NUM	DET	GEN+NUM	GEN+NUM+DET	GEN	NUM	GEN+NUM
Gold Basic	28.62	29.54	29.67	28.41	29.81	29.85	29.91	30.36
+Syntax	28.64	29.51	29.67	28.40	29.86	29.85	29.90	30.38
Trans Basic	28.90	29.55	**29.80**	28.32	29.90	29.91	29.89	30.37
+Syntax	28.87	29.58	**29.80**	28.77	29.90	**30.02**	29.92	30.41
Gold+Trans Basic	**28.96**	29.59	29.77	**28.77**	30.02	29.98	**30.01**	30.42
+Syntax	28.93	**29.60**	29.77	28.75	**30.03**	29.99	**30.01**	**30.43***

Table 6: End-to-end MT results for two translation settings and a variety of added predicted features. Results are in (% BLEU) on our MT05 set. The best result in each column is bolded. The best overall result is marked with *.

maintained (around 3% relative BLEU) as shown in Table 7. Translating through Lemma plus DET and then predicting GEN and NUM is still the best option.

Model	MT05	MT06
Baseline	29.48	19.10
ENGWORD→ARALEM	29.47	18.90
ENGWORD→ARALEM+DET	30.13	19.36
ENGWORD→ARALEM+DET with GEN+NUM Prediction	30.43	19.65

Table 7: Results comparing our baselines and best performing setup on MT05 and MT06 (blind). Results are in (% BLEU).

5.5 Scaling Up

We performed experiments using a larger amount of data (15 times the size of the original dataset; also available from the LDC). Not surprisingly, the effect of our approach diminished. Although the general trends remained the same, none of the alternative settings was able to beat the baseline. We compared the percentage of the Exact Match, Lemma Match and Unmatchable words with the reference of the basic and scaled up systems. We found out that the percentage of exact matches increases while the percentage of unmatched words decreases. This is not a surprising result of using more data. The lemma match percentage decreases across the different systems. This suggests that our approach is more effective for conditions with low and medium resource size.

5.6 Discussion

The generation of fully inflected forms from uninflected lemmas (Table 5) in a purely monolingual setting such as our morphological generation step is very hard – we get only 82.2% BLEU starting with gold lemmas. Adding different combinations of gold values of the three most problematic morphological features improves the score by over 12% absolute BLEU to a higher performance ceiling (94.8% BLEU).

Automatically modeling these features at a high accuracy for SMT, however, turns out to be rather hard. If we consider using them as part of the translation step together with lemmas, we find that they almost always hurt the end-to-end (translation-generation) MT system except for the DET feature which improves over an inflected tokenized baseline by about 0.6% BLEU.

Predicting the feature values using an independent supervised learning step that has access to the English word, POS and syntax features produces accuracy scores ranging in mid to high 80s%. Comparing the prediction accuracy of GEN, NUM and DET (Table 3), we find NUM is the easiest to predict, followed by DET and then GEN. This makes sense given the information provided from English, which is inflected for NUM, but not GEN.

The results in Table 6 show that DET, as a single feature, helps more when it is part of the translation step (30.13 BLEU) compared to being predicted (29.67∼29.80). In both cases, it fares better than simply leaving determining DET to the generation step (29.47).

Neither GEN nor NUM, as single features, help much (or at all) over the baselines when part of the translation step or when predicted. However, when both are combined with DET they consistently help only when GEN and NUM are predicted, not translated. It is possible that the lower performance we see as part of the translation is a product of how we translate: we do not factor these features in the translation – a direction we plan to consider in the future. We postulate that the prediction step helps because it has access to more information than used in our translation step, e.g., source language syntax.

6 Conclusions and Future Work

We compared three methods of modeling morphological features in SMT from English to Arabic: as part of core lexical translation, as part of morphological generation and using an independent morphological prediction component. The best configuration for the three most problematic morphological features for English-Arabic SMT models the determiner as part of core translation and favors predicting gender and number features separately from generation. Our approach shows improvements on a medium-size training data set but when using a very large data set the advantage of using morphological modeling disappears.

In the future, we plan to identify the best configuration for other morphological features in Arabic. We also plan to apply our approach to other target languages such as Persian and Hebrew. We will also investigate how the features we studied here can be used in a more elegant joint model such as Factored MT.

References

Al-Haj, Hassan and Alon Lavie. 2010. The Impact of Arabic Morphological Segmentation on Broad-coverage English-to-Arabic Statistical Machine Translation. In *Proc. of AMTA'10*, Denver, CO.

Alkuhlani, Sarah and Nizar Habash. 2011. A Corpus for Modeling Morpho-Syntactic Agreement in Arabic: Gender, Number and Rationality. In *Proc. of ACL'11*, Portland, OR.

Badr, Ibrahim, Rabih Zbib, and James Glass. 2008. Segmentation for English-to-Arabic Statistical Machine Translation. In *Proc. of ACL'08*, Columbus, OH.

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proc. of EACL'06*, Trento, Italy.

Clifton, Ann and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *Proc. of ACL'11*, Portland, OR.

El Kholy, Ahmed and Nizar Habash. 2010a. Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation. In *Proc. of TALN'10*, Montréal, Canada.

El Kholy, Ahmed and Nizar Habash. 2010b. Techniques for Arabic Morphological Detokenization and Orthographic Denormalization. In *Proc. of LREC'10*, Valletta, Malta.

El Kholy, A. and N. Habash. 2011. Automatic Error Analysis for Morphologically Rich Languages. In *Proc. of MT Summit XIII*, Xiamen, China

El Kholy, Ahmed and Nizar Habash. 2012. Rich Morphology Generation Using Statistical Machine Translation. In *Proc. of INLG'12*, Utica, IL.

Elming, Jakob and Nizar Habash. 2009. Syntactic Reordering for English-Arabic Phrase-Based Machine Translation. In *Proc. of EACL'09*, Athens, Greece.

Habash, Nizar and Owen Rambow. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proc. of ACL'05*, Ann Arbor, MI.

Habash, Nizar and Fatiha Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proc. of NAACL06*, New York, NY.

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

Klein, Dan and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proc. of ACL'03*.

Koehn, Philipp, 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 *Proc. of ACL'07*, Prague, Czech Republic.

Koehn, Philipp. 2004. Statistical significance tests formachine translation evaluation. In *Proc. of EMNLP'04*, Barcelona, Spain.

Lafferty, J., A. McCallum, and F.C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning*.

Maamouri, Mohamed, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *NEMLAR Conference on Arabic Language Resources and Tools*, Cairo, Egypt.

Minkov, Einat, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proc. of ACL'07*, Prague, Czech Republic.

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

Oflazer, Kemal and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *Proc. of ACL'07*.

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

Sarikaya, Ruhi and Yonggang Deng. 2007. Joint morphological-lexical language modeling for machine translation. In *Proc. of NAACL07*, Rochester, NY.

Stolcke, Andreas. 2002. SRILM - an Extensible Language Modeling Toolkit. In *Proc. of ICSLP'02*, Denver, CO.

Toutanova, Kristina, Hisami Suzuki, and Achim Ruopp. 2008. Applying morphology generation models to machine translation. In *Proc. of ACL'08*, Columbus, OH.

Yeniterzi, Reyyan and Kemal Oflazer. 2010. Syntax-to-morphology mapping in factored phrase-based statistical machine translation from english to turkish. In *Proc. of ACL'10*, Uppsala, Sweden.