

# Rich Morphology Generation Using Statistical Machine Translation

Ahmed El Kholly and Nizar Habash

Center for Computational Learning Systems, Columbia University  
475 Riverside Drive New York, NY 10115  
{akholy, habash}@ccls.columbia.edu

## Abstract

We present an approach for generation of morphologically rich languages using statistical machine translation. Given a sequence of lemmas and *any* subset of morphological features, we produce the inflected word forms. Testing on Arabic, a morphologically rich language, our models can reach 92.1% accuracy starting only with lemmas, and 98.9% accuracy if all the gold features are provided.

## 1 Introduction

Many natural language processing (NLP) applications, such as summarization and machine translation (MT), require natural language generation (NLG). Generation for morphologically rich languages, which introduce a lot of challenges for NLP in general, has gained a lot of attention recently, especially in the context of statistical MT (SMT). The common wisdom for handling morphological richness is to reduce the complexity in the internal application models and then generate complex word forms in a final step.

In this paper,<sup>1</sup> we present a SMT-based approach for generation of morphologically rich languages. Given a sequence of lemmas and *any* subset of morphological features, we produce the inflected word forms. The SMT model parameters are derived from a parallel corpus mapping lemmas and morphological features to the inflected word forms.

As a case study, we focus on Arabic, a morphologically rich language. Our models can reach 92.1% accuracy starting only with tokenized lemmas and predicting some features, up from 55.0% accuracy without inflecting the lemmas. If all of the gold morphological features are provided as input, our best model achieves 98.9% accuracy.

<sup>1</sup>This work was funded by a Google research award.

## 2 Related Work

In the context of morphological generation for MT, the state-of-the-art factored machine translation approach models morphology using generation factors in the translation process (Koehn et al., 2007). One of the limitations of factored models is that generation is based on the word level not the phrase level and the context is only captured through a language model. Minkov et al. (2007) and Toutanova et al. (2008) model translation and morphology independently for English-Arabic and English-Russian MT. They use a maximum entropy model to predict inflected word forms directly. Clifton and Sarkar (2011) use a similar approach for English-Finnish MT where they predict morpheme sequences. Unlike both approaches, we generate the word forms from the deeper representation of lemmas and features.

As for using SMT in generation, there are many previous efforts. Wong and Mooney (2007) use SMT methods for tactical NLG. They learn through SMT to map meaning representations to natural language. Quirk et al. (2004) apply SMT tools to generate paraphrases of input sentences in the same language. Both of these efforts target English, a morphologically poor language. Our work is conceptually closer to Wong and Mooney (2007), except that we focus on the question of morphological generation and our approach includes an optional feature prediction component. In a related publication, we integrate our generation model as part of end-to-end English-Arabic SMT (El Kholly and Habash, 2012). In that work, we make use of English features in the Arabic morphology prediction component, e.g., English POS and parse trees.

### 3 Arabic Challenges

Arabic is a morphologically complex language. One aspect of Arabic’s complexity is its orthography which often omits short vowel diacritics. As a result, ambiguity is rampant. Another aspect is the various attachable clitics which include conjunction proclitics, e.g.,  $w+$  ‘and’, particle proclitics, e.g.,  $l+$  ‘to/for’, the definite article  $Al+$  ‘the’, and the class of pronominal enclitics, e.g.,  $hm+$  ‘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). Arabic inflectional features are realized as affixes as well as templatic changes, e.g., *broken plurals*.<sup>2</sup>

These three phenomena, optional diacritics, attachable clitics and the large inflectional space, lead 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 Penn Arabic Treebank (PATB) tokenization scheme (Maamouri et al., 2004), which we use in all our experiments, separates all clitics except for the determiner clitic  $Al+$  (DET). As such we consider the DET as an additional morphological feature.

Arabic has complex morpho-syntactic agreement rules in terms of GEN, NUM and definiteness. Adjectives agree with nouns in GEN and NUM but plural irrational nouns exceptionally take feminine singular adjectives. Moreover, verbs agree with subjects in GEN only in VSO order while they agree in GEN and NUM in SVO order (Alkuhlani and Habash, 2011). The DET in Arabic is used to distinguish different syntactic constructions such as the possessive or adjectival modification. These agreement rules make the generation of correctly inflected forms in context a challenging task.

### 4 Approach

In this section, we discuss our approach in generating Arabic words from Arabic lemmas (LEMMA) using a pipeline of three steps.

1. **(Optional) Morphology Prediction** of linguistic features to inflect LEMMAS.

<sup>2</sup>The Arabic NLP tools we use in this paper do not model all templatic inflectional realizations.

Tokens	$w+$	$s+$	$yktbwn$	$+hA$
POS	conj	fut_part	verb	pron
Lemma	wa	sa	katab	hA
Features	na,na,na, na,na,na, na,na,na,	na,na,na, na,na,na, na,na,na,	3rd,masc,pl, imp,act,ind, na,na,na,	3rd,fem,sg, na,na,na, na,na,na,

Figure 1: An example  $w+s+yktbwn+hA$  ‘and they will write it’. Features’ order of presentation is: PER, GEN, NUM, ASP, VOX, MOD, DET, CAS, and STT. The value ‘na’ is for ‘not-applicable’.

2. **Morphology Generation** of inflected Arabic tokens from LEMMAS and any subset of Arabic linguistic features.
3. **Detokenization** of inflected Arabic tokens into surface Arabic words.

Morphology generation is the main contribution of this paper which in addition to detokenization represents an end-to-end inflection generator. The morphology prediction step is an optional step that complements the whole process by enriching the input of the morphology generation step with one or more predicted morphological features.

We follow numerous previously published efforts on the value of tokenization for Arabic NLP tasks (Badr et al., 2008; El Kholy and Habash, 2010). We use the best performing tokenization scheme (PATB) in machine translation in all our experiments and focus on the question of how to generate Arabic inflected words from LEMMAS and features. Figure 1 shows an example of a tokenized word and its decomposition into a LEMMA and morphological features.

**Morphology Prediction** This optional step takes a sequence of LEMMAS and tries to enrich them by predicting one or more morphological features. It is implemented using a supervised discriminative learning model, namely Conditional Random Fields (CRF) (Lafferty et al., 2001). Table 1 shows the accuracy of the CRF module on a test set of 1000 sentences compared to using the most common feature value baseline. Some features, such as CAS and STT are harder to predict but they also have very low baseline values. GEN, DET and NUM have a moderate prediction accuracy while ASP, PER, VOX and MOD have high prediction accuracy (but also very high baselines). This task is similar to POS tagging

Predicted Feature	Baseline Accuracy%	Prediction Accuracy%
<i>Case</i> (CAS)	42.87	70.39
<i>State</i> (STT)	42.85	76.93
<i>Gender</i> (GEN)	67.42	84.17
<i>Determiner</i> (DET)	59.71	85.41
<i>Number</i> (NUM)	70.61	87.31
<i>Aspect</i> (ASP)	90.38	92.10
<i>Person</i> (PER)	85.71	92.80
<i>Voice</i> (VOX)	90.38	93.70
<i>Mood</i> (MOD)	90.38	<b>93.80</b>

Table 1: Accuracy (%) of feature prediction starting from Arabic lemmas (LEMMA). The second column shows the baseline for prediction using the most common feature value. The third column is the prediction accuracy using CRF.

except that it starts with lemmas as opposed to inflected forms (Habash and Rambow, 2005; Alkuhlani and Habash, 2012). As such, we expect it to perform worse than a comparable POS tagging task. For example, Habash and Rambow (2005) report 98.2% and 98.8% for GEN and NUM, respectively, compared to our 84.2% and 87.3%.

In the context of a specific application, the performance of the prediction could be improved using information other than the context of provided LEMMAS. For example, in MT, source language lexical, syntactic and morphological information could be used in the prediction module (El Kholy and Habash, 2012).

The morphology prediction step produces a lattice with all the possible feature values each having an associated confidence score. We filter out options with very low confidence scores to control the exponential size of the lattice when combining more than one feature. We tried some experiments using only one or two top values but got lower performance. The morphology generation step takes the lattice and decides on the best target inflection.

**Morphology Generation** This step is implemented using a SMT model that translates from a deeper linguistic representation to a surface representation. The model parameters are derived from a parallel corpus mapping LEMMAS plus morphological features to Arabic inflected forms. The model is monotonic and there is neither reordering nor word deletion/addition. We plan to consider these variations in the future. The main advantage of this approach is that it only needs monolingual data which

is abundant.

The morphology generation step can take a sequence of LEMMAS and a subset of morphological features directly or after enriching the sequence with one or more morphological features using the morphology prediction step.

**Detokenization** Since we work on tokenized Arabic, we use a detokenization step which simply stitches the words and clitics together as a post-processing step after morphology generation. We use the best detokenization technique presented by El Kholy and Habash (2010).

## 5 Evaluation

**Evaluation Setup** All of the training data we use is available from the Linguistic Data Consortium (LDC).<sup>3</sup> For SMT training and language modeling (LM), we use 200M words from the Arabic Gigaword corpus (LDC2007T40). We use 5-grams for all LMs implemented using the SRILM toolkit (Stolcke, 2002).

MADA+TOKAN (Habash and Rambow, 2005; Habash et al., 2009) is used to preprocess the Arabic text for generation and language modeling. MADA+TOKAN tokenizes, lemmatizes and selects all morphological features in context.

All generation experiments are conducted using the Moses phrase-based SMT system (Koehn et al., 2007). The decoding weight optimization is done using a set of 300 Arabic 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 4. We report results on the Arabic side of the 2005 NIST MT evaluation set (MT05), our development set. We use the Arabic side of MT06 NIST data set for blind test. We evaluate using BLEU-1 and BLEU-4 (Papineni et al., 2002). BLEU is a precision-based evaluation metric commonly used in MT research. Given the way we define our generation task to exclude reordering and word deletion/addition, BLEU-1 can be interpreted as a measure of word accuracy. BLEU-4 is the geometric mean of unigram, bigram, trigram and 4-gram precision.<sup>4</sup> Since Arabic text

<sup>3</sup><http://www ldc.upenn.edu>

<sup>4</sup>n-gram precision is the number of test n-word sequences that appear in the reference divided by the number of all possible n-word sequences in the test.

is generally written without diacritics, we evaluate on undiacritized text only. In the future, we plan to study generation into diacritized Arabic, a more challenging goal.

**Baseline** We conducted two baseline experiments. First, as a degenerate baseline, we only used detokenization to generate the inflected words from LEMMAS. Second, we used a generation step before detokenization to generate the inflected tokens from LEMMAS. The BLEU-1/BLEU-4 scores of the two baselines on the MT05 set are 55.04/24.51 and 91.54/82.19. We get a significant improvement ( $\sim 35\%$  BLEU-1 &  $\sim 58\%$  BLEU-4) by using the generation step before detokenization.

**Generation with Gold Features** We built several SMT systems translating from LEMMAS plus one or more morphological features to Arabic inflected tokens. We then use the detokenization step to recombine the tokens and produce the surface words.

Table 2 shows the BLEU scores for MT05 set as LEMMAS plus different morphological features and their combinations. We greedily combined the features based on the performance of each feature separately. Features with higher performance are combined first. As expected, the more features are included the better the results. Oddly, when we add the POS to the feature combination, the performance drops. That could be explained by the redundancy in information provided by the POS given all the other features and the added sparsity.

Although STT and MOD features hurt the performance when added incrementally to the feature combination, removing them from the complete feature set led to a drop in performance. We suspect that there are synergies in combining different features. We plan to investigate this point extensively in the future. BLEU scores are very high because the input is golden in terms of word order, lemma choice and features. These scores should be seen as the upper limit of our model’s performance. Most of the errors are detokenization and word form under-specification errors.

**Generation with Predicted Features** We compare results of generation with a variety of predicted features (see Table 3). The results show that using predicted features can help improve the generation quality over the baseline in some cases, e.g.,

Gold Generation Input	BLEU-1 %	BLEU-4 %
LEMMA	91.54	82.19
LEMMA+MOD	91.70	82.44
LEMMA+ASP	92.09	83.26
LEMMA+PER	92.09	83.34
LEMMA+VOX	92.33	83.70
LEMMA+CAS	92.71	84.34
LEMMA+STT	93.92	86.55
LEMMA+DET	93.97	86.62
LEMMA+NUM	93.91	86.89
LEMMA+GEN	94.33	87.32
LEMMA+GEN+NUM	95.67	91.16
++DET	97.88	95.76
++STT	97.73	95.39
++CAS	98.13	96.35
++VOX	98.19	96.47
++PER	98.24	96.59
++ASP	<b>98.85</b>	<b>98.08</b>
++MOD	<b>98.85</b>	98.06
LEMMA + All Features + POS	98.82	98.01

Table 2: Results of generation from gold Arabic lemmas (LEMMA) plus different sets of morphological features. Results are in (BLEU-1 & BLEU-4) on our MT05 set. "++" means the feature is added to the set of features in the previous row.

when the LEMMAS are enriched with CAS, ASP, PER, VOX or MOD features. Our best performer is LEMMA+MOD. Unlike gold features, greedily combining predicted features hurts the performance and the more features added the worse the performance. One explanation is that each feature is predicted independently which may lead to incompatible feature values. In the future, we plan to investigate ways of combining features that could help performance such predicting more than one feature together and filtering out bad feature combinations. The feature prediction accuracy (Table 1) does not always correlate with the generation performance, e.g., CAS has lower accuracy than GEN, but has a relatively higher BLEU score. This may be due to the fact that some features are mostly realized as diacritics (CAS) which are ignored in our evaluation.

**Blind Test Set** To validate our results, we applied different versions of our system to a blind test set (MT06). Our results are as follows (BLEU-1/BLEU-4): detokenization without inflection (55.64/24.92), generation from LEMMAS only (86.70/72.69), generation with gold MOD feature (87.00/73.31), and generation with predicted MOD feature (86.96/73.29). These numbers are generally

Generation Input	BLEU-1%	BLEU-4%
Baseline (LEMMA)	91.54	82.19
LEMMA+GEN	89.23	78.37
LEMMA+NUM	91.14	81.35
LEMMA+STT	91.16	81.70
LEMMA+DET	91.18	81.78
LEMMA+CAS	91.60	82.43
LEMMA+ASP	91.94	83.07
LEMMA+PER	91.97	83.10
LEMMA+VOX	91.99	83.18
LEMMA+MOD	<b>92.05</b>	<b>83.26</b>
LEMMA+MOD+VOX	91.76	82.73
++PER	91.57	82.32
++ASP	91.07	81.32
++CAS	89.71	78.68

Table 3: Results of generation from LEMMA plus different sets of predicted morphological features. Results are in (BLEU-1 & BLEU-4) on our MT05 set. “++” means the feature is added to the set of features in the previous row.

lower than our development set, but the trends and conclusions are consistent.

## 6 Conclusion and Future Work

We present a SMT-based approach to generation of morphologically rich languages. We evaluate our approach under a variety of settings for Arabic. In the future, we plan to improve the quality of feature prediction and test our approach on other languages.

## References

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

Sarah Alkuhlani and Nizar Habash. 2012. Identifying Broken Plurals, Irregular Gender, and Rationality in Arabic Text. In *Proc. of EACL’12*, Avignon, France.

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

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

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

Ahmed El Kholy and Nizar Habash. 2012. Translate, Predict or Generate: Modeling Rich Morphology in

Statistical Machine Translation. In *Proc. of EAMT’12*, Trento, Italy.

Nizar Habash 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.

Nizar Habash, Owen Rambow, and Ryan Roth. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. *Proc. of MEDAR*, Cairo, Egypt.

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

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 *Proc. of ACL’07*, Prague, Czech Republic.

J. Lafferty, 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*.

Mohamed Maamouri, Ann Bies, Tim Buckwalter, and Wigdan Mekki. 2004. The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus. In *Proc. of NEMLAR’04*, Cairo, Egypt.

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

Kishore Papineni, 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.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In Dekang Lin and Dekai Wu, editors, *Proc. of EMNLP’04*, Barcelona, Spain.

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

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

Yuk Wah Wong and Raymond Mooney. 2007. Generation by inverting a semantic parser that uses statistical machine translation. In *Proc. of NAACL’07*, Rochester, NY.